Evaluating the Impact of Physical Activity Apps and Wearables: an Interdisciplinary Investigation of Research Designs and Methods

Claire H. McCallum

Submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy

School of Social and Political Sciences
College of Social Sciences
University of Glasgow

October 2018

Abstract

Many smartphone apps and wearables have been developed to promote physical activity, however there are challenges in assessing their impact. Apps and wearables are rapidly evolving technologies and thousands of physical activity apps that are publicly available on app stores remain unevaluated. There are concerns that traditional "gold standard" evaluation approaches, such as randomised control trials (RCTs), may be too slow to keep up with these, and produce effectiveness results that do not reflect real world settings. Rapid research designs (such as single case designs; SCDs) and innovative data collection methods (in-device sensors, device-generated user logs) have been proposed to improve research efficiency, yet preliminary evidence suggests they are not widely used in mHealth.

This thesis reports three studies undertaken to investigate the use of rapid research designs and efficient methods for evaluating physical activity apps and wearables. First, a scoping review of the extent to which these approaches are employed by health and HCI researchers. Second, semi-structured interviews with researchers, data scientists and industry professionals to provide a deeper understanding of current evaluation practices. Third, the development and refinement of a methodological framework to support researchers in using SCDs in automated app store evaluations of physical activity apps.

The findings suggest rapid research designs are not often employed in evaluations of physical activity and other health behaviour change apps. Researchers feel they face opportunity barriers (e.g. risking not being funded or published) and do not have the necessary skills (e.g. in using device generated user logs). Industry professionals appear to lack the motivation and time to evaluate effectiveness. Trade-offs were perceived between the measurement accuracy of in-device sensors and other factors such as user burden.

Automated trials may speed up evaluations of physical activity apps and wearables, and the suggested data collection framework aims to support researchers in conducting rigorous effectiveness evaluations using app store-based SCDs. However, further work is needed to enable industry professionals to use the framework to evaluate their publicly-available apps.

Table of Contents

List of Tables

List of Textboxes

List of Figures

List of Accompanying Material

Publication of research reported in Study 1: McCallum, C., Rooksby, J., & Gray, C. M. (2018). Evaluating the impact of physical activity apps and wearables: Interdisciplinary review. *JMIR mHealth and uHealth*, *6*(3).

Acknowledgement

I dedicate this thesis to my parents for giving me continuous support of all kinds throughout the entire four years and for helping me in any way they can.

I first must thank my supervisors. Cindy, I really (really) could not have got this submitted without you. I can't quite express how grateful I am for your endless guidance, support and patience, for pulling me out when I got too bogged down in details, for helping me to clarify my thoughts, and providing direction when I needed it. I've learned a lot from you and have really enjoyed working with you.

John, I can't tell you all the ways in which you helped either; but I can say that I'm very grateful for how you have helped me navigate the field of HCI, and provided alternative perspectives on my work, inspiration for the Quped work and reassurance throughout the 4 years. And not judging too much if and when I happen to eat ketchup sandwiches during CHI deadlines.

I am also hugely grateful to Parvin for your help with Quped, Matthew Chalmers for your thoughts on the Quped work, and the rest of SUM group. Matt; thanks for putting up with Worky McWorkface throughout the PhD. I couldn't have done it without your support.

I'm really grateful for my friends, especially Jackie for being able to bounce ideas off you during our Waitrose coffee walks and library sessions, Gözel for your encouragement and openness to letting me share anything with you - both academic and non-academic, and Michelle for your excellent life advice. Beth and Nicola, for always being there when needed, whether it's for a laugh, cry or anything else. Laura, my auntie Jean, Millie, Lindsey, and all others who've provided life support: thank you.

And finally the McCallum clan - I'm very lucky to have you all. Mum, Dad, James, Paul and especially Katie for your wise words and emotional-support-hugs: thank you all for everything.

Author's Declaration

I declare that this thesis is the result of my own work, except where others have contributed as explicitly stated below, and that the work has not been submitted for any other degree at the University of Glasgow or any other institution.

The work presented in relation to Study 1 and Study 3 is the result of collaborations:

Chapter 5 reports the results of Study 1, which was a scoping review. Dr Cindy Gray and Dr John Rooksby contributed to the refinement of the inclusion/exclusion criteria (i.e. the scope of the review), and independently assessed a random sample of studies for inclusion. Dr Cindy Gray and Dr John Rooksby independently extracted data from 5 papers for comparison with the authors' own extraction. Dr Cindy Gray independently coded a random sample of all studies to improve rigor in categorising qualitative data; discrepancies and further discussion were used to inform the final categorisations (i.e. dimensions of engagement, acceptability and usability).

Chapter 9 reports the collaborative aspects of Study 3, which involved designing an app to test and refine a methodological framework developed by the author (presented in chapter 8). The author specified the functionality of the app according to the framework, and Dr Parvin Asadzadeh developed the app to these specifications. Dr Parvin Asadzadeh also designed the user interface, managed the user log architecture, and deployed the app on the Apple App Store. Dr John Rooksby conducted Quped user interviews. The author conducted all data analysis, and managed the project.

Additionally, in Study 2, an undergraduate research assistant helped the author to anonymise transcripts of qualitative interviews.

Definitions/Abbreviations

CEEBIT continuous evaluation of evolving behavioural intervention technology

HCI human computer interaction

mHealth mobile health

MOST multiphase optimisation strategy

OSDAS operationalizing single case designs for physical activity apps distributed via app stores

RCTs randomized controlled trials

SCD single case design

SMART Sequential multiple assignment randomised trial

Chapter 1 Introduction

Many mobile health (mHealth) technologies have been designed for promoting behaviour change, including many smartphone apps and wearables for supporting people to be more physically active. As for any health behaviour change intervention, it is important to know if these technologies can positively impact and increase physical activity levels. However, there is concern over whether traditional "gold standard" approaches to evaluation, such as randomised control trials (RCTs), are sufficient for answering this.

One problem with RCTs is that they can fail to keep up with the rapid rate at which apps and wearables evolve (for example, apps on app stores are continuously updated, or new models of wearables quickly become available). In order to do this, outcome evaluations must become more efficient (Riley et al. 2013, Hekler et al. 2016).

A second issue is that RCTs can be weak regarding their ability to determine whether an intervention would positively impact behaviour in real world settings (i.e. results are often of low external validity). Apps and wearables are typically "complex interventions" and their effectiveness can depend on interactions between their components and real-world contexts (Oakley et al. 2006, Craig et al. 2008). For example, RCTs do not typically focus on real-world engagement and acceptability, yet in the real world settings where apps and wearables are used, engagement can be low, which may negatively influence their effectiveness.

To overcome these issues, new forms of evaluation are needed. In recent years. international researchers have come together within consensus workshops to propose the use of *rapid research designs* as alternatives to RCTs (Kumar et al. 2013, Michie et al. 2017). These include, for example, single case designs (SCDs), the multiphase optimisation strategy (MOST), sequential multiple assignment randomized trial (SMART), continuous evaluation of evolving behavioural intervention technologies (CEEBIT) and microrandomised trials. Furthermore, technological advances mean that efficient data collection methods are now available for assessing not only effectiveness, but also engagement and acceptability. App store platforms can support both efficiency

and external validity by enabling researchers to evaluate apps in the setting they are often made available to the public.

There are a number of reasons why robust evaluations are important. They can: support users and healthcare professionals in deciding which apps to use (Tomlinson et al. 2013); protect consumers from ineffective apps that can often require payment to use (Mohr et al. 2013); reduce opportunity costs (i.e. individuals using apps that are not effective in place of those that are), or users becoming disillusioned with all app-based interventions due to a bad experience with one app (Murray et al. 2016, Michie et al. 2017), and crucially; reduce harm to users (Lewis and Wyatt 2014). For example, physical activity promotion is a popular feature of apps for diabetes (Eng and Lee 2013) and so not understanding their impact could risk user safety.

Given the relevance of rapid research designs and efficient data collection methods for providing insightful evaluations in this area, their uptake and use should be promoted. However, several have noted that rapid research designs, in particular, are not commonly being used to assess mHealth technologies (Blackman et al. 2013, Pham et al. 2016, Jake-Schoffman et al. 2017). This includes low uptake for the evaluation of physical activity apps and wearables. Therefore, a deeper understanding is needed of *why* rapid research designs are not being used for behaviour change technologies. In addition, strategies for increasing their uptake need to be developed.

The overarching aim of this thesis is to investigate the use of rapid research designs and methods for physical activity apps and wearables.

1.1 Research Approach

This thesis employs a mixed-methods exploratory approach, comprised of three studies. Study 1 is a scoping review of research designs and methods used by health and HCI researchers, which was analysed using qualitative thematic analysis and chi-squared testing. Study 2 uses qualitative semi-structured interviews to describe the experiences and perceptions of multiple stakeholders and was analysed using a framework approach (Spencer and Ritchie 2002). In Study 3, a methodological framework was developed and piloted, which involved

reviewing and gathering quality criteria from existing literature, contributing to the design of a physical activity app ("Quped"), and quantatively and qualitatively analysing data from the Quped app store deployment and user interviews.

This thesis is an interdisciplinary investigation, in that it explores the current practices of health and HCI research disciplines in the scoping review and interviews, and Study 3 uses methods employed by HCI researchers (i.e. to design, develop, and deploy the Quped app). Beyond health and HCI, perspectives and experiences across other disciplines and sectors (data scientists and industry professionals) are explored. From a philosophical standpoint, exploring multiple perspectives, interpretations and experiences of a phenomenon provides greater insight into objective truth and reality as it actually occurred (Maxwell and Mittapalli 2010).

While the thesis is primarily focussed on improving the evaluation of physical activity apps and wearables, the work is informed by and can contribute to current debates targeting a range of behaviours. For example, while the scoping review focuses on physical activity apps and wearables, the meaning of these findings are enhanced through interviewing individuals involved in projects relating to a wider range of behaviour change technologies. Similarly, although the framework described in this thesis was developed for use with a physical activity app, it may support researchers in evaluating a range of behaviour change apps, and further testing with other behaviour change apps is encouraged.

1.2 Contributions of this research

Study 1 provides a comprehensive and cross-disciplinary understanding of the state of the field in evaluating physical activity apps and wearables. This includes an assessment of whether rapid research designs and efficient methods, which have been recommended for evaluating these technologies, are actually used.

Study 2 provides further insight into the scoping review results and their significance and relevance to researchers across multiple disciplines. Barriers

that researchers face in using rapid research designs are identified to help explain why they are not being used. Study 2 also provides an understanding of industry professionals' perceptions in relation to evaluating the impact of their publicly available apps. The barriers and facilitating factors identified inform the development of strategies to improve evaluations of physical activity apps and wearables in academia and industry, which are outlined in the discussion chapter.

One of the findings in Study 2 was that academics felt they needed more guidance and support in using rapid research designs. Furthermore, automated studies were identified as an opportunity to increase the efficiency of effectiveness evaluations. Study 3 presents a methodological framework that was developed to support researchers in conducting automated effectiveness evaluations via app stores using one of the rapid research designs recommended for apps and wearables (single case designs, SCDs). The proposed methodology provides a means of evaluating the effectiveness of behaviour change apps that are available on app stores (i.e. within the app store setting in which they are often ultimately distributed), and supporting researchers to use SCDs. The framework was subjected to preliminary testing through its application to the development and trialling of a physical activity app, Quped. The study provides insight into whether SCD studies would be of sufficiently high quality and scientific rigor when conducted via an app store. It also demonstrates how criteria were operationalised in the app to inform the development of future apps aiming to operationalise SCDs. Based on this study, a refined version of the framework is presented that accounts for the natural constraints imposed by conducting a trial in app store settings, whilst promoting rigour.

Drawing on these exploratory studies, the thesis discussion contributes a series of recommendations and suggests avenues for future research in relation to how researchers (and industry professionals) can be supported in evaluating the effectiveness of apps and wearables, efficiently and rigorously.

1.3 Thesis walkthrough

The thesis is comprised of nine chapters. The first four chapters (including chapter 1) are introductory:

Chapter 2 Literature review outlines what is known about physical activity apps and wearables as complex behaviour change interventions (including their features and impact). The chapter then explores challenges in evaluating their effectiveness and identifies associated research gaps. This involves a discussion of: the limitations of RCTs in assessing rapidly evolving technologies; the methodological strengths and limitations of alternative rapid research designs and efficient data collection methods that have been recommended, and their uptake; the limitations of traditional RCTs in relation to external validity and assessing real world impact; and the strengths and weaknesses of app store approaches as a means of improving both efficiency and external validity. Finally, the need for an interdisciplinary approach is explored by outlining what is known about the role of different disciplines and sectors (i.e. HCI, data science and industry professionals) in evaluating physical activity apps and wearables. The chapter concludes with research gaps that were identified.

Chapter 3 Aim and objectives states the overall aim of the thesis and the thesis objectives, drawing on the research gaps identified in the literature review.

Chapter 4 Research Methods outlines the mixed methods approach employed to address the thesis objectives, across three studies. The individual methods used for Study 1, 2 and 3 are described. The chapter then outlines the philosophical approach that was used to support the thesis, how validity has been conceptualised, and identifies limitations of each of the three studies in relation to validity.

Chapter 5 reports the results of Study 1:

Chapter 5 Interdisciplinary review of studies evaluating physical activity apps and wearables presents the findings from a scoping review and provides an initial interpretation of the results.

Chapters 6 and 7 present findings from Study 2:

Chapter 6 Understanding researchers' and industry professionals' perspectives and current practices reports the perspectives and experiences of health and HCI researchers, data scientists, and industry professionals in relation to the key findings from the scoping review (Study 1).

Chapter 7 Encouraging effectiveness evaluations and the use of rapid research designs complements the findings presented in chapter 6 through an-depth analysis (informed by the Capability Opportunity Motivation-Behaviour (Michie et al. 2011) framework) of the barriers and facilitators to evaluating effectiveness and using rapid research designs.

Chapters 8 and 9 report Study 3:

Chapter 8: A framework for Operationalising Single Case Designs for physical activity apps Distributed via App Stores (the OSDAS Framework) presents version 1 (V1) of the three-stage (1. Design, 2. Data Collection and 3. Data Analysis) framework and supporting tools that were developed including an SCD Requirements Checklist and an SCD Quality Analysis Checklist.

Chapter 9: Applying the OSDAS Framework – The Case of Quped reports the application of the OSDAS Framework to the development and app store release of a physical activity app. This involved testing the extent to which the app design, and data collected, could support single case design quality indicators. Drawing on these findings, the chapter presents a refined version of the OSDAS Framework (V2).

Chapter 10: Discussion first presents a summary of the three studies reported in this thesis, and outlines the implications of this work. It provides suggestions on what must change in order to increase the uptake of rapid research designs. Implications in relation to evaluating the impact of publicly available behaviour change apps and wearables, and the methodological strengths and limitations of the three studies are discussed. Finally, recommendations and avenues for future research are provided.

Chapter 2 Literature review of the evaluation of physical activity apps and wearables: what we know so far

2.1 Physical activity: a global health problem

Physical inactivity is the fourth leading risk factor for mortality globally and a major public health problem (Blair 2009), yet physical activity levels are modifiable. Improving physical activity levels can reduce the risk of mortality in those diagnosed with chronic conditions such as cardiovascular disease (Thompson et al. 2003) and diabetes (Church et al. 2004, Sigal et al. 2006), (Chimen et al. 2012), reduce the development of secondary chronic conditions (Lee et al. 2012), and manage and treat obesity which can lead to these chronic diseases (Rippe and Hess 1998, Hill and Wyatt 2005). Physical activity is also highly beneficial for healthy populations without chronic conditions. Observational studies have found that those with increased physical activity have a reduced risk of developing obesity (Reilly et al. 2003) and a range of noncommunicable diseases in later life (Knowler et al. 2002, Shiroma and Lee 2010, Reiner et al. 2013).

Despite the benefits of physical activity in managing and preventing chronic disease, a total of 23% of adults worldwide currently do not meet recommended physical activity levels (35% and 40% in the United States and the United Kingdom, respectively (Mendis 2014). Interventions that have been developed to increase physical activity can be effective (Dzewaltowski et al. 2004), however, traditional interventions delivered face-to-face can be costly. To address the global problem of physical inactivity, further strategies are needed. To this end, there has been significant interest in the use of technology.

2.2 The capabilities of apps, wearables, and app stores in improving physical activity

2.2.1 The functionality and potential reach of apps and wearables

The field of digital health covers a wide range of technologies associated with health and medicine (Lupton 2014). The focus of this thesis, however, is on technologies that can deliver behaviour change interventions and specifically

mobile health (mHealth) technologies. Mobile technologies have been defined as "wireless devices and sensors [including mobile phones] that are intended to be worn, carried, or accessed by the person during normal daily activities" (Kumar et al. 2013, p.228).

While mHealth technologies include mobile devices with limited functionality to deliver interventions (e.g. text messaging services [SMS]), smartphone applications (apps) and body-worn sensing devices (wearables) can deliver highly personalized, individual-level behaviour change interventions. Smartphone apps are supported by platforms (or 'operating systems') such as Apple's iOS or Google's Android (d'Heureuse et al. 2012). Smartphone devices contain a range of in-device sensors (Lathia et al. 2013) such as GPS for tracking geographic location, and accelerometers and pedometers (e.g. Apple's commotion processor), which can collect physical activity data (e.g. steps, and time in moderate-to-vigorous physical activity). Smartphones can connect to wearables through the Internet and Bluetooth technology. Both smartphones and wearables have the computational power to: collect sensor data highly frequently; perform analytics in real-time; and provide feedback to the user on their physical activity levels.

App stores, such as the Apple App and Google Play stores, play an important role in enabling behaviour change apps, including those targeting physical activity, to be publicly available worldwide[1]. Indeed, searching app stores is one of the most common ways for users to find and acquire an app (Nielsen company, 2011). In 2016, there were more than 165,000 health and fitness apps available to users on these app stores (IMS, 2015), with this figure forecast to reach 318,000 in 2017 (Aitken et al. 2017). Many of the apps available have been developed by industry professionals (e.g. product developers and designers) who use app stores to distribute their products to consumers. However, academic researchers who have developed physical activity apps can also ultimately make these available to the public via app stores.

[1] A less common means of distributing apps are app libraries that have been compiled by national public healthcare systems in the UK (https://www.nhs.uk/oneyou/apps) and US (https://www.nlm.nih.gov/mobile/).

Wearables are also available worldwide (Jahns and Houck 2013, Ledger and McCaffrey 2014) and their use is growing (Statista 2017); between 2014 and 2015, wearable sales in the UK grew 118 per cent (Mintel 2016) and were estimated to be a top fitness trend worldwide in 2017-2018 (Thompson 2017).

2.2.2 Apps and wearables are complex behaviour change interventions

Apps and wearables therefore have the potential to greatly increase health behaviour change intervention accessibility and reach (Mechael 2009, Price et al. 2014) and there is growing evidence of their cost-effectiveness (Iribarren et al. 2017). They can encourage individuals to participate in behaviours that protect their health, such as exercising, as well as various other behaviours (e.g. preventing, stopping or reducing drinking or smoking behaviour, Michie et al. 2018).

The theory-based components of the interventions that aim to support behaviour change (i.e. “active ingredients”) are known as Behaviour Change Techniques (BCTs). Michie et al (2011) has developed a taxonomy to categorize these, which includes 93 distinct BCTs. Important BCTs for physical activity include, amongst others, self-monitoring, goal-setting, and social support and social comparison (Michie et al. 2009, 2011). Nevertheless, it has been noted that many app store apps, including those designed to improve physical activity, do not include evidence-based BCTs (Winter et al. 2016).

Digital behaviour change interventions, including physical activity apps and wearables, have been recognized as *complex* interventions (McNamee et al. 2016, Murray et al. 2016). Complex interventions contain multiple components, which can interact with context and produce different outcomes for different people in different settings (Oakley et al. 2006, Craig et al. 2008) and may also be embedded in complex systems (Shiell et al. 2008, Petticrew et al. 2014). McNamee et al (2016) provide illustrative examples in relation to apps and wearables, such as the dynamic nature of social features that incorporate content reliant on other individuals (e.g. social comparison and support), and other complexities that arise when apps are part of a larger intervention partially delivered by healthcare professionals.

The complexity of app interventions is heightened when they are distributed via app stores. Apps can be downloaded from many different contexts and countries worldwide, by people who vary in their characteristics. Users of apps intended for general populations may highly differ in their current physical activity levels and motivation for improving physical activity levels. Overall, apps (including standalone app store apps) and wearables are often complex interventions. This can present challenges for evaluating their impact.

2.3 The impact of apps and wearables on physical activity behaviour

2.3.1 Effectiveness, engagement and acceptability

The impact of an intervention depends on the extent to which it produces effects (Marchand et al. 2011), including any changes in behaviour[2]. Impact is assessed in both efficacy and effectiveness trials, however efficacy trials assess whether an intervention is successful within optimal conditions. This is often within a laboratory or highly controlled setting with minimal complications by other factors. Effectiveness, on the other hand, is the impact of an intervention within "real-world" conditions (Flay et al. 1986), which accounts for the influence of other factors, including whether the intervention is acceptable and actually engaged with.

Understanding real world effectiveness (as opposed to efficacy) is especially important for complex interventions delivered via apps and wearables: user engagement with these devices is typically low (Eysenbach, 2005), which can in turn influence their impact (Donkin et al. 2011, Gilliland et al. 2015). The focus of this thesis is on real world effectiveness, and so the term impact and effectiveness are used interchangeably throughout.

Real world engagement with and the acceptability of complex interventions should be assessed in effectiveness evaluations (Moore et al. 2015) to help interpret and *explain* impact outcomes, i.e. why the intervention worked or did

[2] Flay (1986) more specifically stated a research trial is required to show that the effects are that of "more good than harm" (p.18). Glasgow et al., (2003) notes that many researchers do report harmful effects.

not work in changing behaviour (Oakley et al. 2006, Donkin et al. 2011, Grant et al. 2013, Moore et al. 2015). Digital health researchers have similarly been encouraged to assess 'engagement' and 'acceptability' (Murray et al. 2016), yet how to define and distinguish these constructs has yet to be established. Health researchers often use the term to describe usage behaviour (often focussed on the *amount* of usage) whereas HCI researchers conceptualise engagement as a construct that also includes subjective experiences and perceptions (Perski et al. 2016, Blandford et al. 2018). Perski et al (2016) suggest engagement is a multidimensional construct with both objective (i.e. behavioural) and subjective (i.e. perceptions and experiences) components. Furthermore, in response to varying definitions of engagement, researchers have undertaken valuable consensus-building exercises (Yardley et al. 2016). However, these reviews did not address whether and how engagement should be differentiated from acceptability, which has also been proposed to be a multidimensional construct with objective and subjective elements (Sekhon et al. 2017). Overall for the purpose of this thesis, and whilst acknowledging these to be working definitions, "engagement" is defined as users' interaction and usage behaviour, and "acceptability" as their subjective perceptions and experiences.,

2.3.2 Can apps and wearables improve physical activity levels?

Some systematic reviews have examined the effectiveness of smartphone apps and basic SMS-based interventions for improving physical activity. One early review found 6/11 studies had observed an increase in physical activity (Muntaner et al. 2015), and another reported small increases in activity levels (Blackman et al. 2013), however very few studies were smartphone apps (<4 in both reviews). A more recent meta-analysis of RCTs evaluating mHealth technologies (Direito et al. 2016) included a greater proportion of studies that evaluated smartphone apps or wearables (around half, 10/21). A small effect was found across mHealth technologies targeting physical activity and/or sedentary behaviour. The effect size for physical activity interventions alone was slightly larger (small to moderate), but not statistically significant (that is, there were no differences between control and experimental groups). The authors noted that this might partially be due to the 'active' nature of the control groups studies employed, such as giving participants wearable devices.

Other reviews have included a greater number of studies evaluating physical activity apps and wearables, with mixed results. Stuckey et al (2017) found 8/18 reported increased physical activity levels within intervention groups and 10 studies reported no change in activity levels. Mixed results were also found within a review focused on wearables: Lewis et al (2015) found 5/9 studies reported significant improvements in physical activity. Some reviews have concluded that physical activity apps and wearables can be effective. A recent meta-analysis reported a small effect for apps targeting physical activity, and demonstrated these were more successful in changing behaviour if they includes self-monitoring and goal-setting BCTs (Eckerstorfer et al, in press). Schoeppe et al (2016) found 14/21 studies evaluating physical activity apps reported that these were effective. This review explored whether apps were standalone or embedded in wider behaviour change programmes. Although the latter were more effective the authors noted that "there is still considerable scope to improve the efficacy of app-based interventions" (p. p23).

Overall, with some mixed results, systematic reviews and meta-analyses suggest apps and wearables may have a small effect in improving physical activity, with self-monitoring and goal-setting potentially being effective BCTs. Importantly, even interventions with small effects can be a worthwhile investment if they reach a large number of people, as they can improve overall population health.

2.4 RCTs and the need for efficient alternatives

The above systematic reviews either only included RCTs (Direito et al. 2016), or called for more rigorous research specifically in the form of RCTs (Bravata et al. 2007). While perhaps not surprising, as RCTs are "gold standard" in health research (Haus et al. 2016), what *is* surprising is that these calls overlook an important area of debate surrounding the suitability of RCTs for mHealth technologies (Kumar et al. 2013, Michie et al. 2017). There is a growing awareness that more efficient alternatives are needed to keep pace with the rapidly evolving nature of these technologies.

2.4.1 Randomised control trials

RCTs involve participants being randomly allocated to different groups; either that of the intervention that is hypothesised to have an effect (i.e. treatment or experimental group), or to a "control" group which may, for example, match all other variables except the hypothesised "active ingredient" that is expected to change the individuals' outcome. Hence, any change can be attributed to the difference between the groups (Chambless and Ollendick, 2001). Isolating and controlling variables through this reductionist approach is well established as a means to determine causality (Mook, 1983), and ensure that it is the intervention under study that is responsible for any observed changes in individuals' behaviour.

Smartphone apps and wearables are particularly challenging to evaluate due to the uniquely rapid rate at which they evolve. The design, evaluation and full implementation of traditional behavioural and medical interventions has been estimated to take up to seventeen years to complete (Balas and Boren, 2000) and RCTs themselves can take around seven years to conduct (Ioannidis, 1998). However, smartphone apps and wearables are continuously designed, developed implemented and redesigned, and technologies are quickly superseded (Riley et al. 2013). Entire devices (i.e. hardware) can quickly becoming obsolete, as can apps (i.e. software). Apps are continuously modified by app developers, for example, using "back-end" fixes to ensure apps continue to function as intended and are acceptable to participants, and updates to incorporate entirely new features and functions (Mohr et al. 2015). Furthermore, "just-in-time adaptive interventions" (JITAIS) are *purposefully* built to adapt over time, based on an individuals' data, to deliver an intervention that is continuously more personalised and likely to be effective (Nahum-Shani et al. 2015).

These continuous changes in apps and wearables can directly conflict with the requirement of traditional RCTs for interventions to be stable, static and "locked down" (i.e. unchanged) (Chorpita et al. 2005, Ben-Zeev et al. 2015). Importantly, if the technologies being evaluated become obsolete, this can limit the usefulness of research findings (Mohr et al. 2013, Riley, Glasgow et al. 2013, Patrick et al. 2016). Furthermore, RCTs are expensive to conduct, and if

technologies are no longer available, this can lead to wasted resources (Jake-Schoffman et al. 2017).

Nevertheless, RCTs are still believed to be useful by some mHealth researchers in particular circumstances. Overall RCTs have been deemed suitable when for 'stable' apps that are not rapidly evolving (Murray et al. 2016) for example, are likely to always in be in a state of on-going development and modification in response to on-going user feedback. For such rapidly evolving technologies health evaluations lag behind, and researchers have emphasized the need for greater research efficiency (Kumar et al. 2013, Riley et al. 2013, Hekler et al. 2016, Michie et al. 2017).

Riley and colleagues (Riley et al. 2013) conceptualised efficiency using the "RRR" framework: more *rapid* research (i.e. that is conducted quickly), which is *responsive* to technological changes and advances (i.e. the research accommodates interventions that adapt over time [as opposed to remaining static], and produces results that are *relevant* (i.e. useful for stakeholders, as opposed to, for example, determining the effectiveness of technologies no longer available). Stakeholders might include users and/or healthcare professionals when choosing an app to use from those that are currently available. For physical activity apps and wearables on app stores, this extends to industry professionals who make these apps available.

Hekler and colleagues (Hekler et al. 2016) address the concept of efficiency by proposing that research becomes more "agile". The authors define "agile research" as an "adaptable and nimble scientific process" (p. 317). Thus, it is similar to the "responsive" element of the "RRR" framework (Riley et al. 2013). Agile methods are particularly prominent within start-up app development companies: as they often have few resources, must be highly responsive to changing demands for products and must rapidly (and cheaply) assess the current success of their product (Giardino et al. 2014). This helps them to whether to build on a successful product, or 'pivot' and develop a new product (Bajwa et al. 2017).

Overall, the concept of efficiency can be considered to be speedy research that produces useful results. For results to be useful, however, they must also be

valid. Rapid research designs have been proposed as alternatives to RCTs, which if adhered to, should ensure research is both rapid and rigorous.

2.4.2 Rapid research designs

To increase the efficiency of mHealth evaluations, specific research designs have been recommended which can accommodate rapidly evolving technologies Kumar et al. 2013, Riley et al. 2013, Murray et al. 2016, Michie et al. 2017). Their proposed advantages in increasing efficiency, and their potential disadvantages, will now be addressed in turn.

To evaluate overall effectiveness, the Continuous Evaluation of Evolving Behavioural Intervention Technologies was developed, which tests multiple versions of an app simultaneously (CEEBIT; Mohr et al. 2013). This involves launching a new research trial, and specifically an RCT, each time the app or device is modified. These research trials run at the same time, until one version of the app appears to be less effective, at which point that trial is discarded. The resources required for CEEBIT are likely to be large, and the design is statistically complex. However, while no studies could be found that have used this design, the researchers who conceptualized it propose that a specially designed app store or library could contain multiple app versions to which users can be assigned, or even choose between (Mohr et al. 2013).

To test the effect of individual components of apps and wearables, new rapid factorial approaches have been developed. The Multiphase Optimisation Strategy (MOST) rapidly tests many experimental conditions that isolate different app features (Collins et al. 2005, 2007). MOST allows researchers to understand which components are most effective, and redesign the app accordingly. It is then this optimised 'final version' of the app that can be put forward for testing in an RCT (and as such is complementary to RCTs, rather than a replacement research design), rather than the intervention being continuously changed and optimised throughout the RCT. Other factorial approaches include the Sequential Multiple Assignment Randomised Trials (Murphy, 2005) and Micro-randomised trials (Liao et al. 2016), both of which evaluate components that adapt across time; allowing them to be used to evaluate JITAIS.

Despite their advantages, factorial approaches do have limitations. In MOST, for example, specific, theory-based components of the intervention must be identified in advance that are of interest to test individually (and they may instead be expected to have additive, interactive effects) (Collins et al. 2007). Such decision-making can itself take time and resources (Whittaker et al. 2012). Researchers must also, in advance, assess the feasibility of carrying out the research design, which may include large sample sizes for adequate statistical power (Collins et al. 2007). Furthermore, MOST designs do not necessarily address RCT limitations in relation to external validity and improve understandings of real-world contexts. Nevertheless, these designs are likely a more efficient means of evaluating specific (and potentially time-varying) components, than RCTs.

In addition to the new research designs devised exclusively for digital- and mHealth, "single case designs" (SCD), are an *existing* family of research design that has been generating considerable interest amongst mHealth researchers. There are different types of single case design (i.e. Multiple baseline, AB, reversal, changing criterion and randomized N-of-1s), however all involve participants serving as their own 'control' condition. This requires a baseline phase, and frequent measurement of the outcome. Using the highly frequent, large volumes of data that can be captured by in-device sensors within smartphones and wearables, SCDs can be conducted quickly and easily (Dallery et al. 2013, Hekler et al. 2016). Importantly, unlike new rapid research designs, SCDs benefit from decades of use by researchers in a variety of disciplines such as clinical practice and in education (Guyatt et al. 1990, Smith 2012). Not only have a number of quality standards and checklists accumulated that can be used to ensure any claims or inferences of effectiveness using this design are credible (Horner et al. 2005), but their extensive use mean that, relative to the above rapid research designs, more is known about their strengths and weaknesses.

A central benefit of SCDs above RCTs, and indeed other rapid research designs, is their ability to test the effectiveness of an intervention for a particular individual, in their particular real-world context (Johnston and Johnston 2013, Naughton 2014). Rather than provide a blanket estimation of effectiveness for an "average" individual as in group designs, SCDs can identify *who* an intervention

works/does not work for. They can also be used to assess the effectiveness of individual intervention components (Ward-Horner and Sturmey 2010, Dallery and Raiff 2014). Therefore SCDs can be of utility to HCI researchers who can use results to improve the design of their health apps by including components that work, defining target users, and ultimately, tailoring designs to different users (Klasnja et al. 2011, Hekler et al. 2016).

Although the name "single-case" or "N-of-1" designs suggest that only one individual participates in an entire SCD study, they typically include around six individuals, and there can be greater than 60 (Silverman et al. 1996, Dallery et al. 2013). Multiple SCD trials can be aggregated to produce statistically valid inferences about treatment effectiveness, and developing the methodology for doing so (including multilevel modelling, meta-analysis, and Bayesian inference methods) is an exciting and on-going area of research (e.g. Manolov et al. 2014), (Shadish 2014).

Nevertheless, a widely-acknowledged possible disadvantage of traditional SCDs is the potentially limited generalisability of their results to individuals beyond the research study (Kennedy 1979, Killeen 2018). The ability to generalise from SCDs has been a heated area of debate and fiercely defended. Nock et al. (2007) suggest that SCD studies are not inferior to larger between-subjects designs in their generalisability, as between-subjects designs often employ homogeneous participants in order to isolate causal factors (i.e. the reductionist approach). Similarly, Dallery and colleagues describe the supposed lack of generalisability of SCD studies as a "common misconception" and suggest SCDs are still highly useful and simply involve "carefully choosing the characteristics of the individuals, settings, or other relevant variables in a systematic replication" (Dallery et al. 2013, p.13). Importantly, it was suggested that such replications should be conducted sequentially (i.e. with one experiment taking place after another ends).

Several sequential studies and the decision-making processes involved could take a large amount of time and drastically reduce the efficiency of the research. This would lessen the apparent advantages of SCDs over traditional RCTs. Despite being proponents of SCDs for smartphone-based health behaviour change, Dallery and colleagues (2013) do not acknowledge this issue, nor the

opportunities afforded by technologies to improve generalisability. One approach would be to run automated SCDs using app stores. This approach will be discussed in detail within the later section of this chapter on "RCTs and the need for external validity".

Overall, a shortcoming of RCTs for rapidly evolving interventions is their limited efficiency. Efficiency has been conceptualised in relation to ensuring research is "rapid" (in terms of the speed it is conducted and results are achieved), responsive (accommodating of adapting interventions) and relevant (in producing useful results relating to currently-available technologies) (Riley et al., 2013), as well as "agile" (Hekler et al., 2016). Rapid research designs have been proposed to improve the speed of research. Understanding "what works" may be quicker and more agile with rapid research designs, as they facilitate the assessment of individual app components before or during effectiveness testing (e.g. using MOST, SMART, CEEBIT and microrandomized trials), as opposed to after an RCT (Collins et al. 2007). Yet, it has been reported that the set up of these factorial-based rapid designs, and the associated decision-making required, can be time-consuming (Whittaker et al., 2012). This may reduce their apparent advantages above RCTs. Conversely, It has been suggested that another type of rapid research design, SCDs, are flexible and can be "rapidly implemented", especially when paired with mobile sensors can frequently collect appropriate data in a short period (Riley et al., 1013, p. 3).

2.4.3 Efficient data collection methods

In addition to rapid research designs, the efficiency of research can be improved by using innovative data collection methods. These capitalise on the technological capabilities of consumer devices (i.e. those that are publicly available), which are well-positioned to support evaluations of apps and wearables that target physical activity. In-device sensors that provide feedback to users on their physical activity levels (e.g. accelerometers, gyroscopes and other sensors embedded in smartphones and wearables) can be used to measure outcomes (Dallery et al. 2013, Kumar et al. 2013, Hickey and Freedson 2016). Smartphones can automatically collect steps and activity time (e.g. time spent in moderate-to-vigorous physical activity). Wearables can provide additional data relating to heart rate and calorie expenditure (van Nassau et al. 2016).

Furthermore, internet-connectivity allows sensor data to be transmitted remotely and directly to researchers, allowing them to perform analysis in real time, or store vast quantities of data for later analysis.

The ability of smartphones and wearables to collect continuous, high-density data can improve efficiency over other "intermittent and limited" methods (Kumar et al. 2013), such as self-report questionnaires and pedometers without connectivity. While these "research grade" methods have established measurement validity and reliability, a number of studies have found both wearables (van Nassau et al. 2016) and smartphones such as the iPhone (Major and Alford 2016) to produce valid measures of steps (although less accurate for slow walking speeds).

Although the extent to which health researchers employ in-device sensors to measure physical activity outcomes in impact evaluations is unknown, these sensors have been studied extensively by computing scientists, engineers and data scientists. Lane et al (2010) provide an overview of the 'state of the art' of in-device sensors from a computing science perspective. The authors describe major challenges that arise in managing the large volumes of data collected by in-device sensors.

As well as effectiveness, this chapter has highlighted the importance of exploring engagement and acceptability when evaluating behaviour change apps and wearables. Engagement and acceptability can provide insights into how and why behaviour change apps work. Smartphone and wearables can be programmed to automatically record user interactions and app use (e.g. opening the app, clicks and swipes between screens) (Yardley et al. 2016). Logging methods provide an efficient means of collecting usage data, by reducing the need for users to recall how and when they used the device. HCI researchers have used these device-generated logs to measure engagement objectively (El-Nasr et al. 2013, Dumais et al. 2014), and they are increasingly being proposed as useful for digital health evaluations (Morrison and Doherty 2014, Morrison and Hargood 2014, Sieverink et al. 2016). Although user logs detailing user interaction cannot directly capture the full context and intentions of users in ways that other methods can (e.g. using wearable cameras to support ethnographic approaches (Brown et al. 2013), logs can easily be combined with

other methods (Davies et al. 2017). This includes qualitative methods, to understand why users behaved the way they did (Ploderer et al. 2014, Kwasnicka et al. 2015), and overall acceptability of the technology (El-Nasr et al. 2013), as well as data from in-device 'context sensors' (to understand, for example, geographical location, time of day and the weather (Pejovic and Musolesi 2014) and social situation Lathia et al. 2013).

The use of rapid research designs and efficient data collection methods

The rapid research designs and efficient methods described above may advance mHealth research (Nilsen et al. 2012, Michie et al. 2017), yet evidence suggests that these are not being used by researchers. Pham et al (2016), in a recent review of studies registered in ClinicalTrials.gov, found evaluations of mHealth apps targeting a range of clinical conditions did not use rapid research designs. Instead, the majority of studies used traditional RCTs. Other reviews have also commented on the lack of use of rapid research designs in evaluating apps that are publicly available (Jake-Schoffman et al. 2017), although this statement was not based on empirical research.

One systematic review explored the extent to which studies evaluating physical activity technologies reported internal and external validity (Blackman et al. 2013), and noted need for rapid research designs. However, this was an indirect and 'ad hoc' assessment and most of the studies included in the review evaluated text-messaging/SMS devices, with few included studies assessing smartphone apps (2/15) and wearable devices (2/15). Indeed none of the research conducted to date provides any real insight into the extent to which rapid research designs are used for evaluating sensor-based physical activity apps and wearables. These devices have unique characteristics that can support rapid research designs and efficient data collection methods. Sensor-based technologies can collect data highly frequently, and such continuous 'rich data streams' (Michie et al. 2017) can support the requirements of some rapid research designs such as SCDs (Dallery et al. 2013). It would therefore be useful to understand reasons *why* rapid research designs have so far not been used to evaluate mobile behaviour change technologies. Indeed, there has yet to be a

detailed exploration of the challenges and barriers that may be encountered in using rapid research designs.

Beyond the use of rapid research designs, and in-device sensors, there have been calls for a greater number of studies evaluating behaviour change apps to use device-generated logs and report usage statistics (Schoeppe et al. 2016). Therefore, it would also be of interest to explore the extent to which studies of sensor-based physical activity apps and wearables, specifically, maximise efficiency by using logging software to assess engagement and acceptability

This section has explored the need for efficient alternatives to RCTs. This is particularly important for physical activity apps available on app stores, which are rapidly evolving and unlikely to "stabilise". The next section details what is known so far about the evaluation of these apps and proposes that in addition to efficiency, research designs should promote external validity: namely, whether findings are applicable to and can generalise to other individuals in specific (and real world) settings (Savovic et al 2012).

2.5 RCTs and the need for external validity

Recent reviews suggest that the effectiveness of publicly available apps (i.e. apps available for download on app stores, or "app store apps") is seldom known. A systematic review of physical activity apps available on Apple App Store and Google Play found very few were associated with peer-reviewed publications (Bondaronek et al. 2018). Similarly, two recent systematic reviews that assessed the impact of physical activity apps reported that the majority of included studies had evaluated apps designed and developed for the study, as opposed to those already publicly available (Schoeppe et al. 2016, Stuckey et al. 2017).

While these reviews suggest that few apps developed in industry are subjected to evaluation, there is also the possibility that researchers who do develop (and evaluate) apps do not make these publicly available via app stores. Recently, a review of app evaluations suggested that 72% of studies "had official app names" and used this to infer that these were likely to be intended for distribution to the public (Pham et al. 2016). The authors also found, however, that very few

apps (17%) were publicly available on app stores at the time of the review. This is surprising: as app stores are a central space for users to access apps and have large potential to increase intervention reach (as described above), they would presumably be appealing to many health researchers as a means of disseminating their apps.

Given the particularly rapid and continuously evolving nature of apps, it is especially important that evaluations of apps intended for app store distribution are highly representative of these settings in which they will ultimately be deployed. Glasgow and colleagues (Glasgow et al. 2003) propose that evaluating real world effectiveness from the outset can speed up 'translation' of results regarding the impact of interventions into real world settings. This approach differed to previous evaluation methods which advocated the use of RCTs to first assess impact in highly controlled conditions (Flay 1986).

Traditional RCTs are known to produce results of low external validity, because the trial procedures themselves can influence whether an intervention changes behaviour. Murray and colleagues note that, in the context of digital health technologies, the human support provided by researchers during a trial can artificially boost engagement, whereas in real world settings, app engagement is likely to be low (Eysenbach et al. 2011, Murray et al. 2016). Results from highly controlled studies such as RCTs are particularly unlikely to represent the impact of apps when they are made available via app stores[3]. In RCTs, users are often introduced to the app by researchers and actively incentivized to use it: a user using an app store is presented with thousands of apps that they can browse and select from (Pham et al. 2017).

To increase both efficiency and external validity, an intervention should be evaluated within the settings in which they will ultimately be made available (as opposed to lab-based controlled conditions). Whittaker and colleagues advocate the use of these "pragmatic trials" for evaluating mHealth technologies

[3] This conceptualization of app stores and online marketplaces as real world "settings", contrasts with the earlier presentation of app stores as a distribution method (used by industry professionals and academics to disseminate their apps or interventions to the public. Both are useful conceptualizations and used throughout the thesis interchangeably, where the meaning is hopefully implied by the context and surrounding text. However, considering app stores as *only* a distribution method does not necessarily capture important features of the setting in which app users are introduced to apps, such as being exposed to several competitor apps.

(Whittaker et al. 2012). Similarly, field trials or "in the wild" approaches for studying mobile technologies is an established method used by HCI researchers (Abowd and Mynatt 2000, Rodgers et al. 2015). This involves researchers remotely observing users (with their consent) using mobile devices in their own day-to-day contexts, as opposed to observing them within highly controlled lab conditions. This remote observation is facilitated by user logs. Such passive data collection (as opposed to requiring users to input data) not only improves efficiency but can also improve external validity by reducing the influence of trial procedures on results (such as the Hawthorne effect: an awareness of being measured, Rosenthal, 1966).

In the wild approaches have been used with behaviour change apps and wearables, and even within research that employs an SCD research design e.g. (Kurti and Dallery 2013, Rabbi et al. 2015). Yet in these studies users were still required to come to the lab to meet face-to-face with researchers, for example to be screened for eligibility to participate in the study, to access the app, and to receive instructions for its use. Further, Rabbi et al (2015) provided users with monetary incentives ($120) to promote regular app use over the study period, which may drastically increase engagement. Researcher-participant contact and cash incentives reduce both efficiency (because several face-to-face lab visits can take time and resources to arrange (Volkova et al. 2016) and external validity.

To facilitate studies that are *fully* in the wild (i.e. with little or no researcher-participant contact), researchers can use automation. Riley and colleagues proposed the use of 'automated RCTs', which can evaluate apps while they are being disseminated. However, they did not specify the particular platforms or technologies that would facilitate automation. HCI researchers have also explored the use of app stores (Henze and Boll 2010, McMillan et al. 2010, Morrison et al. 2012) to improve both efficiency (by potentially recruiting and assessing thousands of users with relatively few resource costs to the researcher) and external validity (by reducing biases introduced via contact with experimenters).

Despite the benefits of automated app store based trials in improving research efficiency and external validity, this approach has been adopted by only a

handful of health behaviour change researchers[4]. BinDhim and colleagues conducted an automated RCT to assess a smoking app (BinDhim et al. 2014) and Volkova et al (2016) used the app store to conduct an automated RCT of a nutrition app. These studies required users to find and download the app under study, answer a questionnaire for screening purposes, and be randomly assigned to different app versions. They were thus highly representative of real world effectiveness. To date, one study has combined an app store approach with a rapid research design: Crane et al (2018) used a MOST trial to assess an app targeting alcohol consumption. Furthermore, none of these studies employed in-device sensors to assess the target behaviour, and no studies appear to have employed the app store approach for physical activity.

The app store approach could be particularly useful for SCDs. As mentioned, there are concerns over the generalisability of results produced by SCDs to other individuals and contexts. App stores would enable rapid distribution (and evaluation) of an app to multiple participants in various settings worldwide, whilst collecting data on their characteristics. It would also improve efficiency by eliminating the need to execute SCD studies sequentially, by rapidly conducting several experiments almost simultaneously (i.e. upon launch of the app, multiple users download it). Whether an automated app store approach could support a high quality SCD that produces valid conclusions has not yet been explored.

The limited use of automated app store approaches by health behaviour change researchers means that little is known about the appropriateness or challenges for assessing the impact of physical activity apps. However, HCI studies can provide some insight. Kranz and colleagues report lessons from conducting an experiment on an app distributed via an app store[5]. These lessons included the need for marketing and maintenance of the app in order to acquire many users, and the loss of several users who download the app on incompatible devices

[4] Outside the field of behaviour change, medical researchers have been using app stores in their studies (and industry-based supportive technologies such as Apple's ResearchKit). Currently, however, the objective of many of these studies is to collect observational data regarding health conditions (e.g. to understand diseases). Furthermore, researchers have previously explored automation of evaluations using websites supporting intervention, however this thesis focuses on apps (and wearables).

[5] The app used gamification but was not health related

(Kranz et al. 2013). Henze and Boll (2010), in a study where participants accessed the app on their own devices, found there was wide variability in the mobile phones used. Because the app appearance and associated data collection logging architecture differed across devices, this presented problems in relation to the reliability of results. Morrison et al (2012) also note that determining whether consent to participate is truly "informed consent" (i.e. the participant is aware of the consequences of taking part) can be difficult when using app store approaches. The limited number of studies that have automated effectiveness trials did not report the challenges of automation in depth. Thus whether automating the "MOST" approach differed in difficulty to automating RCTs remains unknown.

Importantly, not all apps will be made available via the app store. A large number may be made available via healthcare settings and supported by healthcare professionals. One framework has recently been developed which supports implementation and impact evaluation in healthcare settings (Mohr et al. 2017). However, the framework is specifically for human-supported interventions, and the authors note that more work is needed to assess the implementation of standalone apps including app store apps. For example, the framework involves gradual withdrawal of the research team, whereas app store approaches include little researcher involvement from the beginning.

Overall, app store approaches may improve external validity over standard RCTS, as they can assess the effectiveness of an app in the context and setting it will ultimately be deployed (i.e. whether the app is found, downloaded and used amongst competitor apps, by real world users of varying characteristics, with little or no researcher contact).

2.6 Evaluating impact: the role of other disciplines and industry professionals

This chapter has so far touched on the importance of incorporating multiple disciplines in evaluations of apps and wearables, including those for physical activity. The importance of multidisciplinary approaches within mHealth research is well established in the wider mHealth literature (Nilsen et al. 2012), and has been recommended by international experts who have come together to

reach consensus on what is needed to improve evaluations (Murray et al. 2016, Michie et al. 2017) recommend use of HCI methods for evaluating acceptability and engagement of digital health technologies, as HCI researchers have considerable understanding of user needs (which is pertinent to understanding acceptability and engagement) and appropriate methods for assessing them. Kumar et al conclude that data science has much to offer in best practices for managing and storing the vast amounts of data that mHealth devices produce, and for detecting patterns within these (Kumar et al. 2013). Michie and colleagues (Michie et al. 2017) note that rapid research designs can make good use of this data.

Despite the above recommendations, very few systematic reviews assessing the impact of physical activity apps have sought to gather evidence from relevant studies from other disciplines. This is particularly important given the usefulness of measuring engagement and acceptability *together* with effectiveness to understand why interventions did and did not work, and to support rapid app redesign and improvement. Unlike health researchers' focus on journals, conferences are the main publication outlet used by HCI researchers to publish high-quality peer-reviewed studies (Blandford et al. 2018). Overall, the narrow focus of systematic reviews means they have likely overlooked, and not taken advantage of, vast amounts of relevant research already conducted within HCI. Only one early systematic review (Bort-Roig et al. 2014), which included only five effectiveness studies of apps and wearables, could be found that included international conference proceedings. These were actively excluded in others (Daskalova et al. 2016, Direito et al. 2016, Schoeppe et al. 2016, Stuckey et al. 2017).

Furthermore, little is actually known about the current practices, perceptions and experiences of those in different disciplines in relation to using rapid research designs and efficient data collection methods. Such an understanding could highlight how efficiency could be further improved, or challenges with rapid and efficient evaluation approaches. As well as HCI, it would be beneficial to gain insight into the practices of data scientists who are familiar with the large data sets produced by apps and wearables. Although consensus workshops with participants from different disciplines have outlined some evaluation

challenges and the need for rapid research designs (Kumar et al. 2013, Michie et al. 2017), consensus-based methods do not necessarily seek to reveal whether and how diverse participants differ in their experiences and perceptions.

In addition to health researchers, HCI researchers and data scientists, industry professionals are particularly important stakeholders in evaluations of physical activity apps. Many of their products are publicly available, yet (as previously discussed) evaluation of their impact is lacking. The involvement of industry professionals in mHealth research has mainly been discussed in the context of their need to adhere to privacy and ethical standards (Tomlinson et al. 2013, Michie et al. 2017), and their perspectives on the innovativeness and disruptiveness of mobile technologies (Whittaker 2012, Sucala et al. 2017). A small pool of literature has touched on the involvement of industry professionals within evaluations studies, specifically. For example, one review discusses the importance of understanding and appreciating the views of software developers in evaluating health systems and technologies (Pagliari 2007). However, more empirical research is needed to explore individual perspectives of industry professionals specifically on mobile physical activity and other health behaviour change technologies.

Conclusions of the literature review

While many physical activity apps and wearables are available and can potentially reach many individuals worldwide, there are challenges evaluating their impact. Traditional RCTs can be appropriate for apps that become ‘stable’ with few further changes, but remain problematic for use with apps that rapidly and continuously evolve (such as those on app stores), or wearables in which new models are frequently released.,

While many physical activity apps and wearables are available and can potentially reach many individuals worldwide, there are challenges evaluating their impact. Traditional RCTs can be appropriate for apps that become ‘stable’ with few further changes, but may be problematic for use with apps that rapidly and continuously evolve (such as those on app stores), or wearables in which new models are frequently released. For rapidly evolving interventions, greater

research efficiency is needed. Rapid research designs have been proposed which can increase efficiency and accommodate continuously evolving apps, however the extent to which these designs have been used to assess the impact of sensor-based physical activity apps and wearables is unknown.

The efficiency of research evaluating the impact of physical activity apps and wearables can be further maximised by using innovative data collection methods, such as in-device sensors. Given that assessing acceptability and engagement in addition to effectiveness can help to provide a better understanding of overall impact, it would be useful not only to know the extent to which all three constructs are assessed by researchers evaluating physical activity apps and wearables, but also the extent to which they are assessed using efficient data collection methods. Current systematic reviews do not incorporate HCI research, which typically addresses engagement and acceptability issues, and thus may be overlooking highly relevant and informative research.

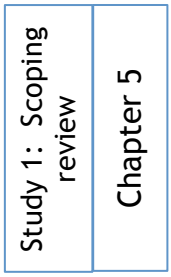
If researchers do not use efficient approaches for evaluating the impact of physical activity apps and wearables, then it is crucial to understand why to illuminate what steps should be taken to improve their uptake. If efficient approaches remain unused, research will continue to be slow and unsuitable for rapidly evolving technologies. Although consensus methods have been used to establish key mHealth evaluation challenges, such as the need for efficiency and rapid research designs, a more detailed examination of diversity in experiences and current practices across disciplines in relation to *using* these designs is needed. This could help to identify any discipline-specific challenges faced, and tailor strategies for encouraging the use of rapid research designs. In addition to health behaviour change and HCI researchers, data scientists and industry professionals are important stakeholders in evaluations of physical activity apps.

Existing RCTs provide little insight into the real world impact of publicly available apps distributed via app stores, which may be the ultimate destination for many of the apps developed in academia. More efficient research approaches that can determine their impact in these real world settings are needed. Combining rapid research designs with automated trials using app store platforms may be a useful approach for standalone physical activity apps. However, no frameworks are available to guide researchers in using this

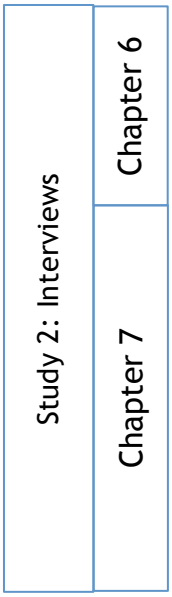
methodology and little is known about its strengths and weaknesses for assessing the impact of physical activity apps and wearables.

Chapter 3 Thesis aim and objectives

The previous chapter reviewed research designs and methods that are available to ensure that effectiveness evaluations keep pace with the rapidly evolving nature of physical activity apps and wearables. Key research gaps were identified in relation to understanding whether and how these are used. The overarching aim of this thesis is to investigate the use of rapid research designs and efficient data collection methods for physical activity apps and wearables. To address this aim there are six related objectives:



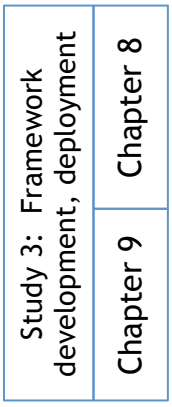
Objective 1. To describe the extent to which evaluations of physical activity apps and wearables: employ rapid research designs; assess engagement and acceptability as well as effectiveness; and use efficient data collection methods



Objective 2. To understand current practices of academic researchers and industry professionals and how they relate to the findings of the scoping review

Objective 3. To identify barriers and facilitators for academic researchers and industry professionals in the evaluation of apps and wearables targeting physical activity and other health behaviours

Objective 4. To identify barriers and facilitators for academic researchers and industry professionals in using rapid research designs



Objective 5. To develop a framework to support academic researchers in using rapid research designs to evaluate physical activity apps distributed via app stores

Objective 6. To test and refine the framework with a physical activity app distributed via an app store

Chapter 4 Research methods

4.1 Introduction

To address the thesis aim and objectives outlined in the previous chapter, three studies were undertaken. These studies employed mixed methods approaches. This chapter will: discuss the philosophical perspectives guiding the thesis; provide an overview of the mixed methods approach; report the individual methods used for each study; outline conceptualisations of validity supporting the thesis; and identify validity concerns for each of the three studies.

4.2 Ontology and epistemology

Ontological perspectives concern "the nature of reality", whereas epistemological perspectives concern theories of knowledge, including "how we know and understand reality". The philosophical stance undertaken in this thesis is that of a 'realist' ontology and 'constructivist' epistemology.

"Realist" ontology holds that there is real, objective truth, independent of our own minds and experiences. This 'mind-independent reality' differs to pure constructivist ontology, which maintains that there is no objective truth; reality is entirely socially constructed (Schwandt 1997, Maxwell and Mittapalli 2010). *Critical* realists advance on realist ontology by proposing a *stratified* ontology. That is, reality has different layers: underneath the surface of what we observe (the empirical) there is what actually occurred (the actual), and responsible for the actual, are complex, underlying causal, generative mechanisms (the real) (Bhaskar 1978). A stratified ontological perspective is taken in this thesis.

A stratified realist ontology allows separation between what is observed, and objective truth enabling a more constructivist epistemology than positivists (Maxwell and Mittapalli 2010). A constructivist epistemology recognises that our empirical observations, experiences and theories will never truly correspond to reality. What we observe will not be the 'objective truth', as our observations are vulnerable to researcher biases and preconceptions, and measurement error.

A stratified realist ontology and constructivist epistemological stance is compatible with using a mixed-method approach (McEvoy and Richards 2006, Maxwell and Mittapalli 2010). Both quantitative and qualitative approaches can be used to observe and describe context and processes. These observations can be used to infer the "actual", and also theorise on possible factors that may be responsible for what we observe (Hartwig, 2015). Thus, measuring observable outcomes (via quantitative methods) is valuable, and interpreting individuals' beliefs, motives, values and contexts (via qualitative methods) is also valuable, to help *explain* these observable outcomes.

Overall, the stance taken in this thesis is that objective truth exists (a realist ontology) but we cannot hope to ever observe or directly measure this truth (a constructivist epistemology). Nevertheless, researchers should strive to maximise the validity of their measurements and interpretations of others cognitive beliefs, motives, values and contexts, and, the validity of the claims, theories and conclusions drawn from these. How validity is conceptualised in this thesis, and its relation to the methods used, is discussed at this end of this chapter.

4.3 Methodology

The research designs and quantitative and qualitative methods used for each study were as follows:

A scoping review was conducted in Study 1, which involved qualitatively coding published articles (i.e. text data) and transforming qualitative codes to generate quantitative descriptive statistics.

In-depth, semi-structured Interviews were used in Study 2. These were conducted via Skype with individuals from different academic disciplines and industry sectors, and were analysed using qualitative thematic analysis.

Framework development and app testing were used in Study 3. This involved collating text from published articles (i.e. gathering requirements) to develop the framework, which was used to inform the

design of a physical activity app. This app was then deployed and tested through an app store, and analysis conducted using both quantitative data from smartphone sensors and user logs and qualitative user interviews.

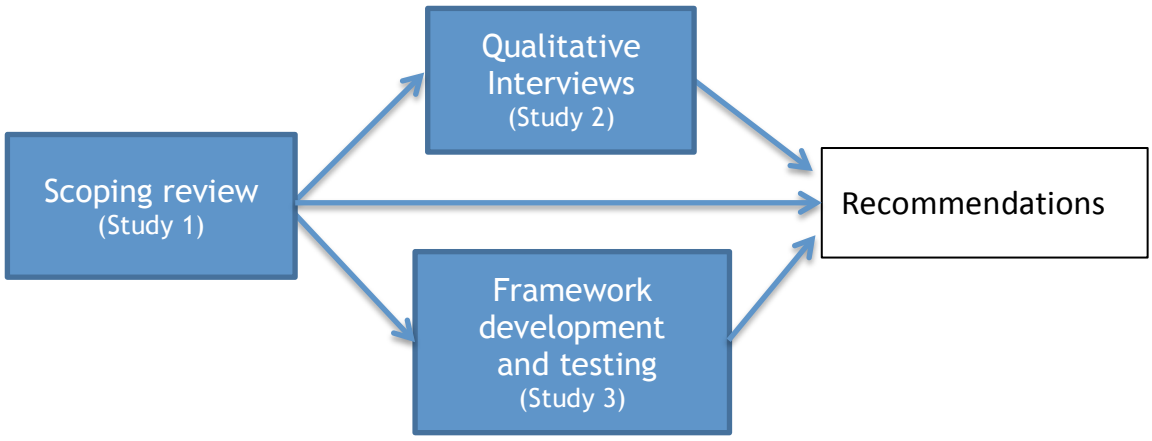


Figure 1: Relationship between the three studies reported in this thesis.

The thesis employs a mixed-methods approach. Specifically, the research reported follows a "dynamic" approach to mixing quantitative and qualitative methods (Creswell and Miller 2000, Maxwell and Loomis 2003), where aspects of existing mixed method typologies are combined, as opposed to following a single mixed methods typology[6] (Creswell and Clark 2007, Creswell et al. 2011). As Figure 1 shows, the methodology consisted of both sequential and concurrent components: the scoping review was followed by two studies running concurrently (qualitative interviews and framework development and testing). Research designs and general methods were established in advance of beginning all studies, based on the thesis objectives. However, emergent findings from Study 1 were used to inform and refine methodological details of Study 2 and 3[7]. All three studies were used to inform recommendations arising from this thesis (as presented in chapter 10).

[6] These typologies describe single 'types' of mixed methods designs, which are characterised by the order in which studies are conducted (sequentially or concurrently), and, the relative weights of quantitative and qualitative research in providing conclusions Creswell, J. W. and V. L. P. Clark (2007). "Designing and conducting mixed methods research.".

[7] Specifically, the interview topic guide developed for Study 2, and the rapid research design used in Study 3, was partially informed by findings from Study 1.

4.4 Study 1: Scoping review

For Study 1, a scoping review was conducted. Scoping reviews are used to rigorously and comprehensively map the range of research activities being undertaken in an emerging field (Arksey and O'Malley 2005), and are thus appropriate for mapping the range of research designs and data collection methods that studies have used to evaluate effectiveness, engagement and acceptability in evaluations of physical activity apps and wearables.

A scoping review was considered to be more suitable than a systematic review. Systematic reviews are typically used to address highly focussed research questions, often surrounding the level of evidence for health care interventions (Armstrong et al. 2011). Systematic reviews therefore assess quality and often reject studies on the basis of study design (and thus favour the inclusion of RCTs) (Kirkevold 1997, Evans and Pearson 2001). This review method would have been counterproductive in light of the need to map a broad range of research designs and methods, and would have also excluded many HCI studies that were likely to use non-randomised designs. Importantly, scoping reviews are *conducted* systematically (through following an established framework), which increases (theoretical) validity and reliability when describing the state of the field (as others could follow the same steps to arrive at similar conclusions). As such, scoping reviews were preferable to narrative reviews. Narrative reviews rely on researchers selecting studies to construct an argument (as opposed to being guided by inclusion and exclusion criteria) (Ferrari 2015); they are therefore more prone to subjective biases and less easily replicated (Yuan and Hunt 2009).

It was important to map the full range of research designs and methods used in order to describe the current state of the field. Searching for specific research designs could also unintentionally exclude or miss studies that used different terminology (which was likely to be the case when including articles from different disciplines).

Scoping review frameworks suggested by (Arksey and O'Malley 2005, and Levac et al. 2010), were adapted to include four steps: 1) identification of relevant articles; 2) study selection; 3) charting and extraction of the data; and 4) collation, summarization and reporting of results (i.e. analysis). A final

recommended step for scoping reviews is to conduct interviews with stakeholders to understand the applicability of the findings to practice (Arksey and O'Malley 2005), as well as act as a knowledge transfer mechanism to enhance the usefulness of the review (Levac et al. 2010). This was conducted as part of the interviews described in Study 2.

Study 1 has been published as McCallum, C., Rooksby, J., & Gray, C. M. (2018). Evaluating the Impact of Physical Activity Apps and Wearables: Interdisciplinary Review. *JMIR mHealth and uHealth*, *6*(3), e58.

4.4.1 Identification of relevant articles

An initial literature search of eight databases was conducted in August-September 2015 and updated in March 2017. Health and clinical databases included PubMed, PsycInfo, and Web of Science, computing science databases included Association for Computing Machinery Digital Library (ACM), Institute of Electrical and Electronics Engineers (IEEE), Springer and Science Direct. mHealth Evidence (a less established but highly interdisciplinary database) was also searched. To maximise sensitivity, keywords were adapted for databases according to discipline (health, computing sciences or interdisciplinary); the search terms used are presented in Textbox 1. Articles were restricted to English language. No time limit was specified. MeSH terms were only used in PubMed, as controlled vocabulary within computing science databases (e.g. ACM's own classification system) was found to return too broad results.

Protocols, conference proceedings and extended abstracts were all eligible. All included articles had been subjected to some form of peer review (extended abstracts are typically peer-reviewed using 'jury' or 'referee' procedures, which are typically used in HCI). As well as those articles returned in the search, the reference lists of systematic reviews were hand-searched for further relevant articles, and if an RCT reported any process evaluation or measures, then the associated protocol was searched for and included in the review to provide further methodological detail.

Databases: PubMed, Web of Science, PsycInfo

Exercise/physical activity/physical activities AND mobile/mobile phone/ smartphone/ sensor/ smart watch/ wearable/wearable device AND intervention/program/app/ application AND evaluate/ evaluation/ assessment/measure/trial/test MeSH terms (PubMed only): “motor activity”, “exercise”, “cellular phones” and “studies with evaluation as topic”.

Databases: ACM, IEEE, Springer, Science Direct

Physical exercise/physical activity/physical activities AND mobile/"mobile phone"/smartphone/sensor/smartwatch/wearable/wearable device/ubiquitous computing AND intervention/program/app/application/activity tracking/personal informatics AND evaluate/evaluation/assessment/measure/trial/test

Database: mHealth Evidence

Physical activity/ physical exercise

Textbox 1: Search terms used in the scoping review

4.4.2 Study selection

Studies were included if they evaluated mobile technologies that provided sensor-based feedback on physical activity. As we aimed to describe the full range of data collection methods used to measure physical activity, studies using objective or self-report measures were both included. Exclusion criteria were: (1) no empirical data was collected (i.e. systematic or methodological reviews, position papers and articles that only described technologies); (2) physical activity was not measured (i.e. studies measured only sedentary time, activity skills, and gait); (3) the study only evaluated sensor/algorithmic performance (i.e. accuracy in recognizing/classifying physical activity); (4) the sensor was not mobile; (5) the only mobile technology used was a pedometer without the capacity to connect to another device or the internet (this exclusion criterion was included in order to focus the review on wearable devices with more advanced feedback capabilities than standard pedometers). All abstracts and full-text articles were reviewed by the author, and 5% of abstracts independently reviewed by supervisors. Discrepancies were discussed between the author and supervisors, and all were resolved.

PRISMA guidelines, which outline how to conduct and report systematic style reviews (Liberati et al. 2009), recommend merging multiple publications on the same study to ensure research is not over-represented. In the current review,

where more than one article referred to the same study, these were merged to represent one study.

4.4.3 Data Extraction

A data extraction form was developed to reflect the objectives of the study. The form was piloted on three articles, revised, and agreed upon with supervisors before being applied to all included studies. Items for extraction included 1) study characteristics (i.e. publication year, country of study, number of participants, age of participants, study duration, whether a protocol or full trial); (2) research design details (i.e. experimental/non-experimental design, number of groups, experimental/control group details, randomisation) and intervention characteristics (i.e. technologies/devices used to deliver intervention, key intervention features); (3) research objectives and outcomes measured; (4) analyses undertaken (descriptive, inferential, thematic); (5) data collection methods used (e.g. in-device or external sensors, user-logs, questionnaires, interviews, focus groups). To promote consistency and reliability of data extraction, the author and two supervisors (CG and JR) extracted five papers (5%) independently.

4.4.4 Collation, summarization and reporting of results (i.e. analysis)

A mixed-methods descriptive approach was used to analyse the extracted data (Levac et al. 2010). First, frequencies were calculated for each research design identified. The intervention characteristics (i.e. components or app features that studies evaluated) were also mapped. Next, the research objectives and outcomes that studies measured, as reported by authors, were used to categorize studies according to whether they investigated effectiveness (i.e. changes in physical activity). Categorising studies according to whether they investigated engagement and acceptability required a more iterative approach, as definitions of these constructs are less widely agreed. Working definitions of engagement (i.e. user interaction with the device and usage behaviour) and acceptability (i.e. users' subjective perceptions and experiences), were used: these were applied to extracted research objectives, outcome measures and data collection methods to develop a series of broad codes in relation to

engagement (i.e. engagement, usage, use, adherence, compliance) and acceptability (i.e. acceptability, satisfaction, user experience, usability). These codes were applied to all studies to allow them to be categorised according to whether they investigated engagement and/or acceptability. Frequencies are reported for the number of studies in each category.

In relation to effectiveness, the proportions of studies that: used only descriptive statistics (as opposed to inferential statistical analysis) was calculated. Studies that used sensors were grouped by whether they used in-device sensors, and/or b) external sensors (i.e. additional, validated devices) to collect physical activity data, and frequencies were then calculated for the data collection methods used in each group. A Chi-square test of independence was conducted to examine if the type of sensor used was related to the type of research design using R statistical software (RStudio, version 1.0.136).

In relation to engagement and acceptability, the data collection method extracts were firstly used to calculate frequencies in relation to the data collection methods employed (e.g. user-logs, questionnaires, focus groups, interviews). Each extract was then read carefully to identify detailed sub-codes that described the different elements assessed for each construct (i.e. any specific behaviours logged, questionnaire items used, or interview/focus group topics described), and the One Sheet of Paper method (Ziebland and McPherson 2006) used to generate broad dimensions of engagement and acceptability by grouping these sub-codes according to their similarity.

A random sample of all studies (20%, 23/111) was independently coded (by one supervisor, CG) to improve rigour in categorising studies and generating the dimensions in relation to engagement and acceptability; discrepancies were discussed and consensus was reached on the final dimensions. Discussions with supervisors suggested that some of the dimensions initially associated with acceptability were specifically related to the properties of the app or device, and therefore did not relate to acceptability per se. These dimensions were retained and recategorised as ‘usability’.

4.5 Study 2: Interviews with researchers and industry professionals

4.5.1 Research Design

For Study 2, 15 semi-structured interviews were conducted using Skype. Participants were recruited from academia (health behaviour change researchers, human computer interaction researchers) and industry (industry data scientists, CEOs, product designers) to represent a range of perspectives. Semi-structured interviews were chosen, as there were particular questions that the author wished to address. They were considered more appropriate than fully structured interviews (i.e. where closed questions are delivered in a particular sequence requiring restricted answers) or questionnaires. Open-ended questions enabled wider issues to emerge beyond the main 'evaluation' topic chosen by the author to enrich analysis; participants could voice opinions on other important issues that could then be immediately followed up (or probed) for clarification (Barriball and While 1994). Furthermore, in advance of interviewing people from across different disciplines and sectors (i.e. academia and industry), it was not always clear what to expect (e.g. the degree to which respondents conducted any evaluations of behaviour change apps). Semi-structured interviews allowed the author to adapt questions to ensure they were relevant to the interviewee (or at least, the depth at which to pursue particular questions), modify the question order to enhance the interview flow, and build rapport with each individual participant. Semi-structured interviews also allowed the author to follow up and check any misunderstandings or misinterpretations of terminologies belonging to particular disciplines (e.g. HCI and health professionals may have different conceptualisations of 'effectiveness').

Semi-structured interviews were also preferred over group-based approaches, including focus groups and Delphi methods. Focus groups can clarify views (Kitzinger 1995) and generate consensus and convergence (Sim 1998); homogenous focus groups (i.e. consisting of participants from the same discipline/sector) could have generated key issues for each discipline/sector, which could have been compared across groups to reveal differences. However, there were several reasons for not using focus groups. Practically, although focus

groups can be conducted remotely (Davis 2001, Hennink 2013), participants were geographically diverse, and so arranging a time that suited everyone (across time zones) would have been difficult. Furthermore, a focus group consisting of industry professionals from different companies could have generated competitiveness or reluctance to share business secrets. These group dynamics were not of interest for the research objectives, and could interfere with findings. Finally, focus groups can lead to 'surface-level' discussion (Powell and Single 1996); interviews allowed in-depth exploration of a wide range of experiences.

Another group-based alternative to semi-structured interviews was a Delphi study. This would have involved gathering data individually from a selection of experts, aggregating these, and reporting the results back to participants who would then adjust their responses to allow some form of consensus to occur (Adler and Ziglio 1996). A previous study has already sought consensus on the challenges of evaluating mHealth technologies for behaviour change using Delphi methods (Michie et al. 2017). It was of far greater interest in the current study to explore *differences and diversity* in perspectives and experiences between (and within) disciplines, and between (and within) industry and academia. Furthermore, a Delphi approach would have required greater time commitment from participants and restricted the time available to explore individuals' backgrounds and contexts in-depth (which was considered key to understanding the complexity of challenges faced).

4.5.2 Participant sampling and recruitment

A purposive sampling strategy (Anselm and Corbin 1998, McEvoy and Richards 2006) was used for Study 2. Participants were recruited based on the author's perception that they had relevant expertise and experiences, and could provide insight into the beliefs held by *some* members of that discipline in particular contexts. As there was no desire to make generalisations to the discipline as a whole (i.e. a wider population), random sampling was not used.

Participants of interest were those who had designed and/or evaluated a technology targeting health behaviour change (including, but not limited to, physical activity). Projects involving medical/diagnostic technologies were not

of interest. Participants belonged to three relevant academic disciplines involved in mobile behaviour change technology projects (as suggested by literature in this area (Kumar et al. 2013, Michie et al. 2017): health behaviour change, HCI, or data science. Both senior and early-career researchers were of interest. Industry professionals were of interest because they were expected to reveal challenges in using rapid research designs and efficient methods beyond academia, and also allow any other 'real world' issues associated with evaluating health behaviour change apps and wearables to emerge. Industry professionals could have any position in a company (CEOs through to designers and developers).

A sampling frame was devised to initially include 12 academic researchers (four health, four HCI and four data science), and four industry professionals. However, identifying individuals as academic 'data scientists' was found to be difficult as it turned out to be a less 'discrete' discipline area; instead, individuals appeared to belong to a specific discipline and simply apply data science techniques within that discipline. This lead to the sampling frame being revised throughout the interview process.

The majority of Health and HCI academics were recruited from conferences attended by the author. Therefore, the sampling strategy contained an element of 'convenience' sampling, as well as purposive sampling (Marshall 1996). The author engaged in informal discussions with these individuals and identified them as having relevant expertise, then sent a formal follow-up email invitation to participate in an interview. One academic participant was recommended by the author's supervisor.

Data scientists and industry professionals required more varied recruitment strategies: the author became aware of one data scientist having come across their published article, and another data scientist at a conference. Two industry professionals were also encountered at conferences. Three industry professionals had released successful apps on app stores: one was CEO of a company who had developed a leading physical activity app and was sent a direct email invitation; to recruit the two others, the author contacted an employee of a technology company to ask if she knew of anyone who may be interested in participating. This employee then approached two industry professionals (from other

companies) and, after gaining their consent, introduced them to the author via email.

Interviewees were broadly categorised according to the sector in which they worked at the time of interview (i.e. academia or industry). Two interviewees worked in both sectors and were categorised according to which of these the author perceived was their predominant place of work, and one interviewee was categorised in academia despite also having a governmental role. Although data scientists worked in industry, they were categorised as data scientists to distinguish them from other industry professionals. Overall, participants (n=15) were eight academics (4 health behaviour change, 4 HCI), five industry professionals (3 CEOs, 2 product designers) and two data scientists. They worked in different countries including the UK (7), USA (2), France (1), Australia (1), Netherlands (1), Germany (1), Sweden (1), and Israel (1).

4.5.3 Interview procedure

A topic schedule was devised based on the research objectives and scoping review findings. Items included: participants' background (i.e. disciplines and experience in academia/industry); the research designs and methods participants used; their familiarity with rapid research designs and efficient data collection methods (including conducting research remotely and at scale); the challenges experienced in conducting evaluation research; and what was needed to advance evaluation research. Participants were also asked for their views on key findings that emerged from the scoping review, and were encouraged to share any relevant experiences. Bryman and Cassell (2006) have noted that when recruited for their expertise and knowledge in a particular area, academic researchers may feel pressure to answer "correctly". Therefore, where possible, interview questions were deliberately worded to elicit opinion, as opposed to factual answers.

After devising the topic schedule, pilot interviews were performed with one academic (in HCI) and one industry professional. These transcripts were reviewed with supervisors and revised before conducting any further interviews. Specifically, some questions were reworded to try and access people's experiences, rather than their speculations. It was also decided that the topic

schedule contained too many questions: questions addressing interviewees opinions on conducting studies remotely and at scale were excluded, as the pilot interviewees had little experience of these methods. Data from the pilot interviews were included in the final analysis.

Most interviews were around 45 minutes, ranging from 30 minutes to 1.5 hours. All interviews were audio recorded with participant consent and transcribed by a professional company approved by the University of Glasgow. During the interviews, field notes were jotted down which mostly included either words and phrases the author expected would not be clear in the audio recording (due to a poor Skype signal), or, key points and ideas that would be later elaborated on within theoretical notes generated immediately after the interview (Altrichter and Holly 2005). Generating theoretical notes involved summarising the key issues mentioned by the interviewee, critically reflecting on the interview as a whole, and considering possible connections to the experiences reported in other individuals' interviews. These notes were typed up electronically, continuously revisited to add further reflections as they occurred (Elo and Kyngäs 2008, Phillippi and Lauderdale 2017) and used to inform the development of the thematic framework (see section 4.5.6).

4.5.4 Framework Analysis

Transcripts were anonymised and each participant given a unique pseudonym for use within the thesis. A framework approach was used to analyse the interview data. The framework method (Spencer and Ritchie 2002) is a form of thematic analysis that can be used to not only describe attitudes and perceptions but also illuminate possible *explanations* of social behaviour (i.e. address diagnostic questions which 'examine the reasons or causes of what exists'). This approach was considered useful for describing the interviewees' experiences and perceptions and beyond this, understanding why they were not using rapid research designs and methods. Unlike other qualitative methods such as grounded theory, the framework method permits themes to be rigorously examined both across participants, and in detail for individual cases (Spencer and Ritchie 2002). This was necessary for exploring differences in experiences and perceptions between broader participant categories (i.e. 'researchers' and 'industry professionals'), and also sub-categories within these (e.g. different

research disciplines and positions in industry). The framework process involved: familiarisation; identifying a thematic framework; indexing; charting; and mapping and interpretation (Spencer and Ritchie 2002, Gale et al. 2013).

4.5.5 Familiarisation

Transcripts were read as a whole once the vast majority of data had been collected. Potential themes that were recorded in theoretical notes during the data collection process (i.e. hunches, Spencer and Ritchie 2002) were explored across the wider data set. The range of participant responses to each interview question, and the different issues that participants thought were important (such as explicit 'challenges' when evaluating mHealth), were noted.

4.5.6 Creating a thematic framework

Creating a thematic framework is akin to the 'coding' stage in other qualitative analysis methods. Six early transcripts were selected for diversity of participants and their circumstances[8]. Based on the research objectives and potential themes identified in the familiarisation stage, these transcripts were used by the author to generate nine initial broad codes. The transcripts were also given to two supervisors (three to CG, three to JR) to independently generate broad codes. The broad codes where then discussed in a face-to-face meeting, where some of the authors' original broad codes were collapsed and others, considered less relevant to the research question, excluded. The resultant five broad codes were: Personal and organisational context; Projects; Research Design; Assessment; and Future/what's needed.

4.5.7 Indexing

The first round of indexing involved applying the above broad codes to all transcripts, using NVivo 10. Then, each broad code was reviewed, and using a bottom-up approach, several emergent *sub-codes* were generated. These sub-codes were then refined (i.e. similar sub-codes were collapsed) to create a final set of sub-codes associated with each broad code. Then, in the second round of

[8] Transcripts were those of two health researchers (one early career researcher, one in a senior position), two industry professionals (one who evaluated behaviour change, and one who did not), and two HCI researchers

indexing, sub-codes were systematically applied to all transcripts, again using NVivo 10. Multiple codes could be applied to a single text item (to allow ideas to emerge for associations between codes in the later interpretation stage).

4.5.8 Charting

A chart or 'matrix' was created for each broad code using Microsoft Excel, containing sub-codes as columns, and cases (i.e. individual participants) as rows. Cases were grouped by discipline (Health, HCI, data science) and sector (academic or industry). Charting involved summarising coded data and entering it into the appropriate cell in the chart, whilst keeping the 'essence' and meaning of what was said by the participant (often using participants' own phrases).

The framework method can involve iteratively revisiting different stages of analysis (Spencer and Ritchie 2002) and revising their outcomes. The charting process, which provides the opportunity to view all content associated with each code, led to the further refinement and re-indexing of sub-codes (i.e. revision of the thematic framework). As the author had applied multiple codes to text items when indexing, this meant that the same text appeared in more than one chart or column. Instances of this were closely examined, and, after deciding which codes the text was most closely associated with, it was then often assigned to a single code (i.e. put in a single cell). In some instances, text was kept in multiple cells to ensure that important associations between codes were retained: at this point, the author began to engage in the next Framework stage (mapping and interpretation). Overall, iteration between the charting and interpretation stages lead to some sub-codes being associated with different broad codes, and the 'Project' broad code was removed; relevant text within transcripts was re-coded in NVivo 10, and four charts were created reflecting the final four broad codes and their sub-codes (which are provided in Appendix 1).

4.5.9 Mapping and Interpretation

After charting the data into broad codes and sub-codes, a deductive approach was used to address the three research objectives in Study 2.

To address objective 2 (which was to understand current practices of academic researchers and industry professionals and how they relate to the findings of the scoping review), the broad codes and sub-codes perceived to relate to the key scoping review findings were reviewed. For example, for the scoping review finding that most included studies used RCTs, data in the chart associated with relevant broad codes (research design) and sub-codes (experience using RCTs, perceptions/knowledge of RCTs) were reviewed across all participants. Interviewees' perceptions and experiences were clustered according to their similarity, and any differences were noted. These similarities and differences were then interpreted in relation to participants' sector and discipline (to contrast these).

During the above process, some findings emerged that did not directly map to the key scoping review findings. These mostly consisted of research designs and data collection methods that participants currently used in their everyday practice but were not captured by the scoping review. These were assigned to a separate sub-code (i.e. Further Research Designs and Methods).

To address objectives 3 and 4, the COM-B model was applied to the charted broad codes and sub-codes. The COM-B model is used to analyse Capability, Opportunity, and Motivational factors that influence a particular Behaviour (Michie et al. 2011). Michie and colleagues proposed that these factors may help to explain a particular behaviour, and why people may *not* engage in that behaviour, and as such can inform the development of strategies for behaviour change (Michie et al. 2011). As shown in Figure 2 the factors or "components" can interact and inform behaviour directly or indirectly (for example, Opportunity and Capability factors can influence an individuals' Motivation and Behaviour), and each component has subcomponents. Michie et al (2011) describe Capability as "individual's psychological and physical capacity to engage in the activity concerned" (Michie et al. 2011, p.4), while Opportunity is influences on the behaviour that are external to an individual, comprising of physical factors (including time and resources) as well as social opportunities (e.g. having supportive cultural or social context). Motivation is "'all those brain processes that energise and direct behaviour" including reflective motivation, i.e. "evaluations and plans", "goals and conscious decision-making" and

"analytical decision-making" (Michie et al., 2011, p.4). Motivation also includes automation motivation, such as emotional responses.

COM-B has often been used to characterise barriers and facilitators for people engaging in health behaviours (such as improving physical activity levels or eating habits) (e.g. Webb et al., 2016). However, COM-B can also be used to characterise professionals' behaviours, such as prescribing or delivering interventions (Michie et al., 2011). The target behaviours in Study 2 related to the evaluation practices of professionals (i.e. researchers and industry professionals, as opposed to the health behaviours of app users). Specifically, two separate behaviours were analysed: the evaluation of impact (objective 3), and the use of rapid research designs (objective 4).

The COM-B model was chosen over other theoretical frameworks. Normalisation Process Theory (NPT), in particular, was a credible alternative. NPT focuses on healthcare professional behaviours (May, 2009) (as opposed to "patient" behaviours), and also explores how behaviours can be implemented within professionals' routines (which is similar to exploring how evaluation behaviours and rapid research designs can be implemented within researchers' routines). However, beyond understanding barriers to using and implementing specifically rapid research designs (objective 4), Study 2 explored the wider aim of barriers in evaluating effectiveness more broadly (objective 3). Although COM-B is a broad and simple model, it is flexible (Barker et al., 2016) and was considered useful in its applicability to both objectives 3 and 4. Furthermore, COM-B is a component of the wider "Behaviour Change Wheel" which maps different types of barrier to different theoretically derived solutions (Michie et al. 2011). As such, using COM-B in this research allows researchers in future to explore relevant potential strategies that address any barriers identified.

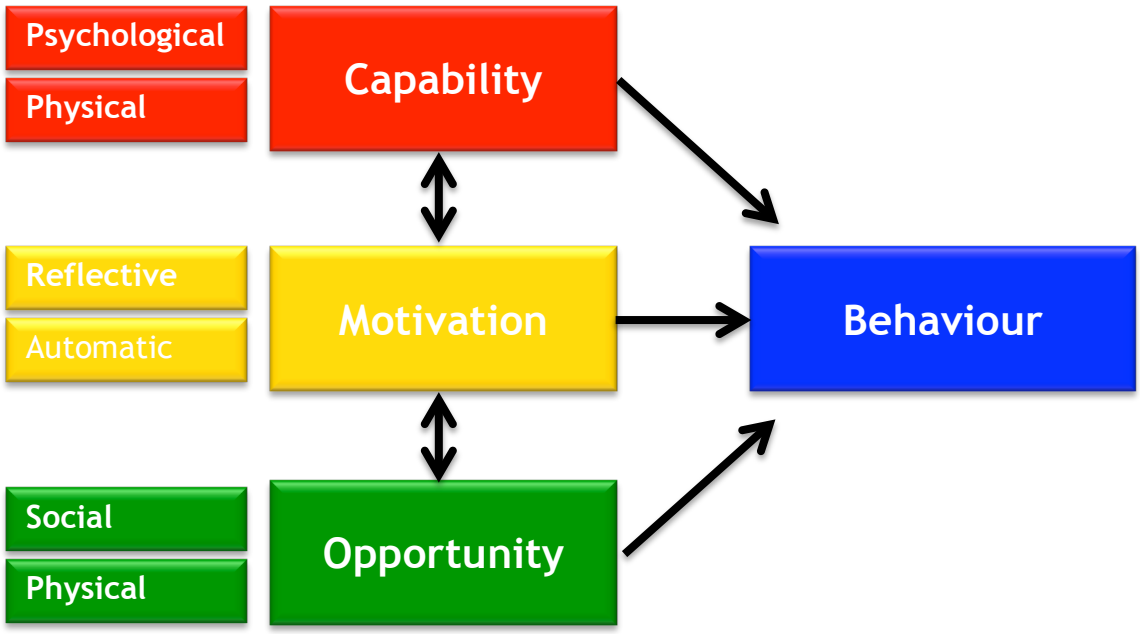


Figure 2: The COM-B Model
Comprises of Capability, Motivation and Opportunity factors that influence a Behaviour. Adapted from Michie et al. (2011) to display sub-components (left).

Analyses to address objectives 3 and 4 were conducted concurrently. After reviewing all charted sub-codes (independently of their broad codes), those perceived to be irrelevant to objectives 3 and 4 (primarily those relating to data collection methods) were excluded from this part of the analysis. Then, data associated with each sub-code were categorised according to whether they related specifically to "using rapid research designs" or "general evaluation". Data within each category were clustered (i.e. across sub-codes), according to perceived similarities in the perceptions, experiences and concepts that participants described, mapped to individual dimensions of the COM-B model (i.e. psychological capability, physical capability, social opportunity, physical opportunity, reflective motivation, and automatic motivation) (Michie et al. 2011)[9], and any barriers or facilitators of evaluation of impact and the use of rapid research designs identified. Differences across participants were noted, along with their sector and discipline.

[9] No data were found to be associated with two COM-B dimensions (i.e. physical capability and automatic motivation), and so these were subsequently removed from analysis.

4.6 Study 3: Framework development and testing

The final study (Study 3) was the development of a Framework to support researchers in conducting rapid research design studies with physical activity apps distributed via app stores. The Framework focuses on single case designs (SCDs), and Operationalizing SCDs within physical activity apps distributed via App stores (The OSDAS Framework). Importantly, the framework is not intended to support researchers in analysing effectiveness analysis itself; rather, it focuses on the construction of a research design that can support the validity of any effectiveness claims. This section outlines the methods used to develop the OSDAS Framework and its components, and to subsequently test and refine these components.

4.6.1 Framework development

HCI literature describing app store methods (Henze and Boll 2010, Kranz et al. 2013, McMillan et al. 2010, Morrison et al. 2012, Weber et al. 2016) were reviewed to conceptualise three basic stages of the OSDAS Framework: Stage 1) designing an app and its logging architecture; Stage 2) deploying the app on the app store to collect log data remotely; and Stage 3) analysis of logged data. To tailor the app store approach to support an SCD study, existing SCD criteria within established standards and guidelines were identified and collated to form two methodological “checklists” for use in the OSDAS Framework. Specifically, checklists were designed to support Stage 1 (by providing a list of SCD requirements that should be operationalised in the app’s design and logging architecture) and Stage 3 (by providing a list of questions that should be answered during data analysis) of the Framework.

Overall, the OSDAS Framework development process was iterative: for example, one of the SCD criteria identified (“social validity”) required qualitative data to demonstrate the criteria had been met. Therefore, Stage 2 of the Framework (specifically, stage 2) was modified to incorporate qualitative interviews alongside and app store launch, a methodology originally proposed by HCI researchers (Morrison et al. 2012).

4.6.1.1 Identifying SCD criteria

To identify SCD requirements, exploratory literature searches were conducted using Google Scholar, and citation chaining (i.e. following which articles cited an article, or hand-searching that article's reference list) (Webster and Watson 2002, Jalali and Wohlin 2012). A structured systematic review of specific databases was not considered appropriate; initial literature searches indicated that there was wide variety in the disciplines using SCDs and the terminology employed (which may be omitted in a systematic review with pre-specified search terms). Exploratory searches returned numerous checklists, recommendations, standards, guidelines, and proposed best practices for conducting, reporting, and reviewing SCD studies. The final set used to inform the framework included: the What Works Clearinghouse Standards (Kratochwill et al. 2013); the Single Case Experimental Design (SCED) criteria (Tate et al. 2008); Risk of Bias in N-of-1 Trials (ROBINT) (Tate et al. 2013), the APA Division 16 Task Force on Evidence-Based Interventions in School Psychology (Kratochwill et al 2003), and those described by Horner et al. (2005), Dallery et al. (2013) and Klein et al. (2017). These articles contained broad criteria (e.g. the study must promote *internal validity*), with specific methodological indicators of whether criteria are met (i.e. "quality indicators").

4.6.1.2 Developing framework checklists

As there is not one set of widely agreed upon standards for SCDs (Smith 2012), the criteria and quality indicators included in the above sources were collated to identify a single set of criteria and their associated quality indicators for use in the OSDAS Framework. This involved first reviewing, familiarizing and noting differences in the above standards and checklists. Then, all quality indicators were gathered across standards and checklists. Quality indicators that related to evaluating effectiveness were excluded from the development of the OSDAS Framework (n=3), as evaluating effectiveness was beyond its scope. Next, quality indicators that were perceived to be highly similar were merged. Finally, the remaining unique quality indicators were firstly grouped into methodological criteria categories (e.g. "internal validity"). to form the "SCD Requirements Checklist" component of the OSDAS Framework. Secondly, quality indicators that required data analysis (e.g. to determine whether data showed "stability") were

identified and grouped together to form a separate checklist: the “SCD Analysis Checklist” component of the OSDAS Framework.

4.6.2 Testing the OSDAS Framework

Testing the OSDAS Framework involved using it to design, deploy and analyse a physical activity app store app, and assessing the extent to which the app design and data collected met SCD quality indicators and criteria. This testing process was used to optimise the OSDAS Framework checklists for future app store deployments.

4.6.2.1 Design and development of a physical activity app

It was considered necessary to invest the time and resources in developing a new physical activity app and releasing it on the App store, as opposed to using an existing physical activity app that had been previously developed and released by industry professionals. Using an existing app would have significantly reduced the time and resources required to develop, market/advertise and maintain the app, and more advertisement could have led to greater participant numbers and a greater volume of data. However, several issues were anticipated with this approach, including: limited access to the data collected (as it would be mediated by an app development company); restricted control over how the intervention was delivered and experimental conditions manipulated; contamination issues whereby participants were already exposed to or familiar with the app; and ethical issues, whereby ‘trying out’ an experimental design may have influenced the success of the app and company profits. As such, the ‘Quped’ app was developed, which was intended to represent a ‘typical’ physical activity app store app.

4.6.2.1.1 Behaviour Change Techniques

Content analyses (Bondaronek et al., 2018, Direito et al. 2014, Middelweerd et al. 2014) ranking the behaviour change techniques most commonly implemented in physical activity apps on app stores, were used to design Quped. Top-ranking techniques included self-monitoring and feedback, and goal-setting (Bondaronek et al., 2018, Direito et al., 2014, Middelweerd et al. 2014). To inform Quped’s goal-setting algorithm (i.e. the increments by which goals automatically

increased each week), studies of traditional pedometer interventions (e.g. Tudor-Locke and Bassett 2004, Tudor-Locke, Sisson et al. 2005) were consulted. HCI research (which explored user preferences for different ways in which behaviour change techniques could be implemented in a PA app) indicated that users preferred to begin weekly goal-setting on a Monday, as opposed to the day of first use (Munson and Consolvo 2012). Social comparison was another behaviour change technique often employed by physical activity app store apps (Bondaronek et al., 2018, Middelweerd et al. 2014). The social comparison feature also facilitated collection of demographic data; app store policy requires that data should only be collected on users if it is beneficial for them (i.e. by entering details, users are able to compare their steps with those of the same or different demographics).

4.6.2.1.2 App store trial features

To understand which app features were necessary to facilitate an automated 'fully remote' trial with little researcher input, the author reviewed studies that had used app store approaches (or methodological literature which discussed it) within health behaviour change (BinDhim et al.2014, Volkova et al. 2016), and within HCI (Morrison et al. 2012, Weber et al. 2016). Together, studies had included features to support informed consent, questionnaire screening for participant inclusion in the study, and tutorials on how to use the app. While some questionnaires only required users to select their age and gender (Weber et al. 2016), others included more detailed questionnaires (Volkova et al. 2016). In the Quped trial, a detailed screening questionnaire was not needed as there was no specific population of interest. Therefore, the Quped app was designed to simply include: remote trial features (including participant information) to support informed consent and withdrawal; and tutorial screens upon launching the app.

4.6.2.1.3 Operationalizing SCD requirements

The author consulted with computing science researchers (PA & supervisor JR) to explore how to design and implement (i.e. 'operationalise') the OSDAS Framework SCD requirements checklist (i.e. SCD criteria and quality indicators) in an app store deployment of Quped. Consultations involved a series of

prototyping (i.e. transforming ideas into tangible early app versions to develop and test ideas (Walker et al. 2002) and focussed discussions. First, a series of app screen sketches (i.e. low-fidelity prototypes) that supported behaviour change techniques and remote trial features were generated. Then the author led focussed discussions with the computing science researchers on how to operationalise SCD criteria within the app. These discussions centred on technical feasibility (i.e. how and when screens could be introduced to users, and what data could be logged), Apple app store policies (e.g. restrictions on collecting unique data for a single participant, and the need for data collection to be useful for the participant), user behaviours required to facilitate the design, and the likelihood of maintaining user engagement (and thus obtaining data required).

Final app screens, how and when these screens would be introduced to users, and what data to log were agreed, and final sketches were coded up and implemented (by PA) into a fully-functioning app (i.e. a high-fidelity prototype). Before releasing the app on the App Store, the app was tested internally[10]. Bugs were fixed before the app was launched on the App Store, and internal test data were removed from the final dataset before analysis.

4.6.2.2 Data collection: Launching the app on the app store and user interviews

4.6.2.2.1 Releasing and testing the App store app

The Quped app was released on the Apple app store in February 2016. A sample of data, consisting of the first 6 months of deployment (i.e. 27 weeks from February to August 2016) was used in this study. To understand the extent to which data collected from an app store release met SCD criteria, measurable indicators of success were generated before the app was released. These were based on how quality indicators in the SCD Requirements Checklist had been operationalised in Quped, including when the screens were expected to be delivered, how users were expected to behave, and how the data-logging architecture was intended to function.

[10] Internal testing involved installing the app on the research teams' iPhones, using it for two weeks, and then checking data logs to reveal any 'bugs' (i.e. whether the intervention components were delivered, and whether data was retrieved, as intended).

4.6.2.2.2 Recruitment

To participate in the trial, individuals were required to have an iPhone 5S or iPhone 6, as the Quped app would not work on earlier versions of the iPhone or on Android devices. In addition to simply releasing the app on the App Store for people to discover by chance, the recruitment strategy used included targeted social media advertisements (via Facebook) to people aged 18+ who used an iPhone 5S or above and advertisements placed in online university newsletters. Both advertisements contained hyperlinks directing users to the App Store. App features enabled users to view participant information and provide in-app consent to take part in the study. If users did not consent, or later withdrew from the study (via an in-app menu that was continuously available), they could continue to use the app; however, any data that had already been collected was excluded from analysis, and no further data was logged.

4.6.2.2.3 Measures

Google analytics software was used to create an estimate of how many users downloaded the app. The majority of SCD quality indicators were tested using in-app log data. Log data involved a combination of a unique identifier for the individual, step data, app interaction data, time stamp data and information (see Table 1).

4.6.2.2.4 Interviews

The final SCD criterion, social validity, was assessed using qualitative semi-structured telephone interviews that were conducted by supervisor JR primarily to explore the acceptability of the Quped app. However, the author worked closely with JR to ensure the topic schedule contained questions relating to social validity, and led the related analysis.

Data collected	Data logged
Unique identifier for each download	Apple does not permit identification of individuals, therefore we could only identify individuals by each download of the app (a resulting limitation being that a user could theoretically install, delete, and reinstall the app [or use it on two devices] and would be counted in the study as two individual users).
Step counts	Data retrieved from the internal pedometer (M7 chip) or

	native Health Kit app for each 24-hour period (i.e. integers associated with a time stamp).
Interaction data	Whether users interacted with the app on that day (true/false); whether users registered for the social comparison feature; the age bracket and gender (M/F) that users selected to allow social comparison
Goal automatically generated for each user	The goal set in integer form. On days where no goals were set (i.e. the baseline phase and intervention phase 1 "B"), goal = 0
Timestamp and timezone data	Date and time for every day that user step count data were collected (including the dates for the six days that baseline phase data were retrieved). Although time-zone data were provided for each daily step value collected, users were categorised as belonging to one time-zone only (i.e. the time-zone with highest frequency for that user)

Table 1: Log data collected during the app store trial

Recruitment to the interviews was conducted separately from the app store trial. Social media messages and posters (around the University campus) were distributed, and interested individuals (either already using Quped, or those interested in downloading and using it for the purposes of the interview) contacted JR, who then interviewed participants approximately four weeks after they had downloaded the app. Incentives (£10 in cash) were given to participants who only consent to use Quped for the interview study (i.e. not existing Quped app store users) for agreeing to run the app on their phone and to be interviewed. Eighteen participants took part in interviews, which were transcribed and anonymised; including 13 who downloaded Quped during the first six months of the app store release and five who were recruited after the six month release and asked to download Quped (in order to further explore social validity).

4.6.2.3 Data analysis

Log data collected during the first 6 months of deployment of the Quped app were analysed using Microsoft Excel and R software[11]. Notably, as the aim of the

[11] A computing scientist retrieved data from logging software and sent this to the author (in a CSV format) for analysis.

study was not to understand effectiveness of the Quped app, there was no hypotheses testing of effectiveness, and no inferential statistics were used.

4.6.2.3.1 Visual analysis

The author produced visualisations using the R packages “ggplot 2”, and “SCVA” (Bulté and Onghena 2012) to explore user step patterns and support the use of descriptive statistics, as explained below. A visualisation showing users downloading the app at different times was also created to assess internal validity criteria. This facilitated ‘eyeballing’ (but not quantifying) the extent to which overlap/verification periods occurred.

4.6.2.3.2 Statistical analysis

Descriptive statistics (counts, and counts as a percentage of all consenting users) were calculated for the number of participants who met different criteria. Baseline stability was assessed by calculating relative variability around the mean (Schoenfeld et al. 1956, Costa and Cançado 2012, Blackman 2017)).

4.6.2.3.3 Qualitative data analysis

Specific quality indicators associated with social validity from the SCD analysis checklist, along with an initial familiarisation of the interviews, were used to create the following five codes: step count importance; procedure acceptability/intrusiveness; perceived app effectiveness; intent to continue use; and data privacy. These codes were applied to all transcripts. Then, data within each code were grouped based on the similarity of topics described to reveal the dimensions of each broad code. These dimensions were then interpreted in relation to original quality indicators.

4.7 Ethics

Ethical approval was sought from the College of Social Sciences at the University of Glasgow for the interviews conducted in Study 2 (400150139). The study was considered to be generally low risk to participants, as the interview topics were not believed to be sensitive in nature. No incentive was offered to participants. It was appreciated that participants may wish to be acknowledged and

recognised for their contributions. However, in order to maintain anonymity as far as possible, participants were informed that they would not be acknowledged within any published work directly relating to the interviews.

Ethical approval was acquired for the Study 3 Quped feasibility trial and associated user interviews from the University of Glasgow, College of Social Sciences (400150014). The Quped app was designed to support informed consent and users could withdraw at any time via an app menu.

4.8 Conceptualising and promoting validity

The stance taken in this thesis is that validity is a property of a claim (i.e. conclusion, theory or inference), rather than a method itself, and so is not tied to either a quantitative or qualitative approach (Cook et al. 1979, Shadish et al. 2002). There are different types of validity, depending on the type of claim or conclusion a researcher makes (Maxwell 1992, Johnson 1997). Of particular relevance to the research reported in thesis is "interpretive" and "theoretical validity". This section will outline: these different types of validity and how they apply to the quantitative and/or qualitative methods used in this thesis, possible threats to that validity and methods that were used to address these to promote validity.

4.8.1 Interpretive validity

Interpretative validity concerns the ability to describe participants' own perspectives: whether a researcher's report of participants' beliefs, meaning and interpretations reflects, as closely as possible, what participants indeed feel (Maxwell 1992, Hammersley and Atkinson 1995). Interpretative claims involve interpreting 'mental' or cognitive objects (e.g. cognitions, beliefs), and as such interpretive validity does not apply to quantitative methods (Maxwell 1992).

Within the interviews used in Study 2, interpretive validity was strengthened during the framework analyses, by using 'low inference descriptors' (Johnson, 1997): often including participants' own verbal phrases when summarising data, to preserve meaning (Spencer and Ritchie 2002). A potential threat to interpretive validity relevant to both Study 2 and the interview analysis in Study

3 was that a deductive approach was incorporated to address the research objectives. Deductive approaches are often used in applied research to answer specific questions (Spencer and Ritchie 2002), but risk skewing participants' interpretations to 'fit' the research question. To counter this, the interviews analyses in both Studies 2 and 3 involved the explicit step of familiarisation with the data to ensure that broad codes were an appropriate fit for all data. In Study 2, inductive coding was used within broad codes and sub-codes to interpret these according to the prior scoping review findings (objective 2) and the COM-B framework (objectives 3 and 4). Furthermore, in Study 2, a broad code (Personal and organisational context) included interpreting context from the accounts of participants, which aided the process of writing from their 'point of view'.

4.8.2 Theoretical validity (including internal and external validity)

Theoretical validity involves the greatest degree of 'abstraction' or inference from the data, and is the degree to which researchers are justified in the conclusions they draw. These theories, inferences and conclusions can involve either: proposing the existence of certain constructs or concepts, or, generating 'causal' explanations (Maxwell 1992). A specific type of theoretical validity relevant to the former can be considered 'construct validity', which applies to both quantitative and qualitative research (Adcock, 2001). In the literature review of this thesis, it is acknowledged that defining the constructs of engagement and acceptability is an on-going area of debate, and so working definitions are employed. Some of the conclusions drawn from Study 1 rest upon these working definitions. Nevertheless, the theoretical validity of these conclusions was otherwise promoted by: using multiple data points in the form of multiple articles ("data triangulation", Denzin 1970); discussing constructs with supervisor CG (i.e. "peer debriefing" (Johnson, 1997); and incorporating independent coding where CG coded 20% of relevant articles for comparison, discrepancies were discussed and codes revised to better reflect the definitions of engagement and acceptability used (i.e. investigator triangulation, Denzin, 1970).

4.8.2.1 Internal validity

Internal validity is the extent to which a researcher is justified in making a causal claim or theory (Cook et al. 1979). It is the view of the author that while qualitative data can be useful to generate theories and ideas on causal mechanisms and factors in particular contexts, such theories are highly 'fallible', and should be continuously revisited and revised. Thus, as discussed in Chapter 10, the barriers and facilitators proposed to influence behaviour (e.g. whether rapid research designs are used) in Study 2 should ultimately go on to be tested experimentally within particular contexts, in future studies. Nevertheless, internal validity was strengthened by interviewing multiple people from different sectors and multiple disciplines, to gain multiple perspectives on reality.

The number of interviewees in Study 2 (15 overall, with fewer in each sub-category e.g. "four HCI researchers") means that full "data saturation", where more participants would have provided few additional insights, is unlikely to have been reached. However, it has been suggested by Malterud and colleagues (2016) that "saturation" is a concept based on grounded theory (Glaser and Strauss, 1999), an approach that was not employed in Study 2. Malterud et al., 2016 propose a different and more relevant approach to assessing sample size in qualitative research: namely the use of "information power". The authors suggest that fewer participants are needed if the sample offers substantial information power (and conversely, that smaller information power requires larger participant numbers). Factors proposed to influence information power include (i) the study aim (ii) specificity of the sample (iii) use of theory (iv) quality of the dialogue, and (v) analysis strategy.

Notably, Malterud et al., 2016 describe information power "as an aspect of internal validity" (p. 7). Using the information power approach, and guided by the above factors, the sample size used in Study 2 required fewer participants to maintain the internal validity of conclusions. In relation to (i), the objectives (2, 3 and 4) were fairly narrow (i.e. participants' views on specific scoping review findings, their use of particular research designs, their views relating to behaviour change app evaluations); (ii) the sample included researchers and industry professionals who were specifically involved in the development and/or

evaluation of mobile behaviour change apps, (iii) a pre-existing theory, COM-B, was applied to achieve objectives 3 and 4, (iv) participants were fairly articulate about their experiences and many understood the concepts of evaluation, therefore fewer participants were needed to provide rich and relevant data, and (v) framework analysis was used which employs "case level" analysis, as well as cross-case. This meant that, although some conclusions were made using small participant sub-groups (i.e. 3 "HCI" researchers", 4 "health" researchers"), individual participants' personal and organisational contexts were explored in detail.

Neither Study 1 (the scoping review) nor Study 3 (developing the OSDAS Framework) make causal claims. Nevertheless, the OSDAS Framework proposed in Study 3 is used to support causal claims, and follows the logic that to increase internal validity, experiments should be designed to increase robustness against alternative, rival causal explanations (i.e. 'validity threats', such as maturation and history effects) (Shadish et al. 2002, Johnson 1997, Maxwell and Mittapalli 2010). Notably, a typical, established means of improving internal validity in experimental paradigms is randomisation (i.e. randomly assign participants to different experimental conditions to overcome participant level biases (Shadish et al. 2002). However, while incorporating some form of randomisation within SCDs is possible, SCDs do not typically require randomisation to meet internal validity criteria (Dallery et al. 2013) Instead, SCDs employ "replication logic" (Yin, 1984) whereby replicating the experiment with similar individuals increases confidence in the causal relationship (i.e. internal validity of claims that the app had an impact on particular participants). SCD logic aligns with the philosophical stance undertaken in this thesis, whereby no inferences are made to a 'general population', but only to specific, similar contexts.

4.8.2.2 External validity

The final form of validity that should be judged for any (causal or non-causal) theories proposed is external validity, or generalisability. This is the extent to which theories generated from observable data apply to contexts, settings, or participants that have not been observed. This applies to both qualitative and quantitative data. The philosophical stance in this thesis previously outlined (i.e. stratified realist ontology and constructivist epistemology) supports the

view that is possible to create abstractions or models of a process that occur in a particular context, and then use these models to understand other comparable contexts (Ragin, 1987). This 'process' theory of causation focuses on context-dependency, rather than generalising across entire populations and different contexts, but nevertheless permits some generalisation in the form of *demi-regularities* ('patterns that hold imperfectly over a restricted region of space-time' (Hartwig 2015, p116). The framework qualitative analysis method used in Study 2 allows close and rigorous examination of cases, as well as codes (Spencer and Ritchie 2002); i.e., during all analyses, the author paid close attention to participants' individual contexts. By reporting these contexts in the results, other researchers can assess the relevance of the findings of Study 2 to other contexts in future work.

Employing reflexivity, where the researcher explicates their own influence on qualitative data collection and analysis, can improve a study's external validity. Interviews are themselves social situations and there is always a participant-interviewee relationship that could bias results and reduce certainty that the same conclusions would be drawn if other participants were included, or another researcher conducted the study. The author had prior contact with 9/15 participants in Study 2, and thus an opportunity to build rapport. As a result, participants may have provided more open and honest answers, thereby reducing any social desirability biases. However, a limitation is that some interviewees had seen the author present a conference paper on single case designs, which may have influenced participants' awareness of these designs. Other individuals, had they been selected for interview, may have different levels of awareness. Accessing participants in industry required a more varied recruitment strategy with less prior contact, which may have accounted for some of the differences in attitudes between academia and industry.

Reflexivity also helps to improve the reliability of conclusions drawn. Within qualitative research, reliability can be considered to be the extent to which other researchers would, if following the processes and analytical procedures outlined, arrive at similar conclusions (Robson and McCartan 2016). Throughout the work reported in this thesis, documentation of methodological decisions and processes (i.e. audit tracing) increases reliability by allowing those external to

the project to see how the author arrived at conclusions and therefore replicate the study (Schwandt and Halpern 1988). The scoping review in Study 1 and framework approach in Study 2 both followed systematic, established processes (Spencer and Ritchie 2002, Levac et al. 2010) that outline each step taken to arrive at the conclusions. While the author did have 'hunches' about issues considered likely to emerge from the interviews in Study 2 (from her own experiences as a researcher, and having previously interpreted scoping review findings), the framework method ensures that all aspects of transcripts are given equal attention (Spencer and Ritchie 2002), to ensure theories are tied to data. In Study 3, the specific SCD standards and checklists that were used to arrive at a set of quality criteria and indicators for use in the OSDAS Framework are explicitly stated.

4.9 Summary

This chapter has first outlined the author's ontological and epistemological perspectives. It has then described the quantitative and qualitative approaches employed across the three studies in this thesis (the scoping review, interviews and framework development and testing), and the reasons for these methods being chosen over others. The chapter concluded by exploring some of the validity concerns for each study and describing the methods used to address them. These validity concerns, and others which emerge, will be further discussed after the results are presented in the following three chapters.

Chapter 5 Evaluating the Impact of Physical Activity Apps and Wearables: An Interdisciplinary Scoping Review

5.1 Introduction

This chapter present findings from Study 1. The study was devised to address objective 1, which was to describe the extent to which evaluations of physical activity apps and wearables: use recommended rapid research designs; assess engagement and acceptability as well as effectiveness; and employ efficient data collection methods (i.e. in-device sensors and device-generated logs).

The discussion presented at the end of chapter provides an initial interpretation of results. It also concludes with suggestions for further research that contributed to the design of studies 2 and 3. Findings from the current chapter therefore inform the research reported in chapters 6-9.

The work in this chapter has been published: McCallum, C., Rooksby, J., & Gray, C. M. (2018). Evaluating the Impact of Physical Activity Apps and Wearables: Interdisciplinary Review. *JMIR mHealth and uHealth, 6*(3), e58. http://doi.org/10.2196/mhealth.9054

5.2 Results

5.2.1 Summary of search results

A total of 6,521 articles were retrieved during the initial database search (see Figure 3). After title screening, 1,272 abstracts were reviewed and 645 articles not meeting inclusion criteria were excluded. The full texts of the remaining 627 articles and an additional 12 articles identified from reference lists searches were read. 572 studies were excluded, leaving 68 articles.

An additional 56 articles were included from the updated search in March 2017. Here, 557 abstracts were reviewed and 338 articles not meeting inclusion criteria excluded. When the 219 remaining full texts were read, a further 163 articles that did not meet criteria were excluded. Therefore, a total of 124 articles were included in the review, representing 111 unique studies.

5.2.2 Study characteristics

The study characteristics are presented in Table 2. Of the 111 included studies, 22 (19.8%) were protocols. Over half (61/111, 55.0%) were published in 2015 or later. Many (47/111, 42.3%) were conducted in the USA. The majority of studies (103/111, 93.0%,) involved adult participants; eight studies (7.0%) involved children and adolescents. Participant numbers ranged from 2 **(**Albaina et al. 2009) to 2,980 **(**Gomes et al. 2012**)**: 19% (21/111) of studies contained fewer than 13 participants. Study duration ranged from less than a day to 52 weeks. Intervention characteristics are included in Appendix 2.

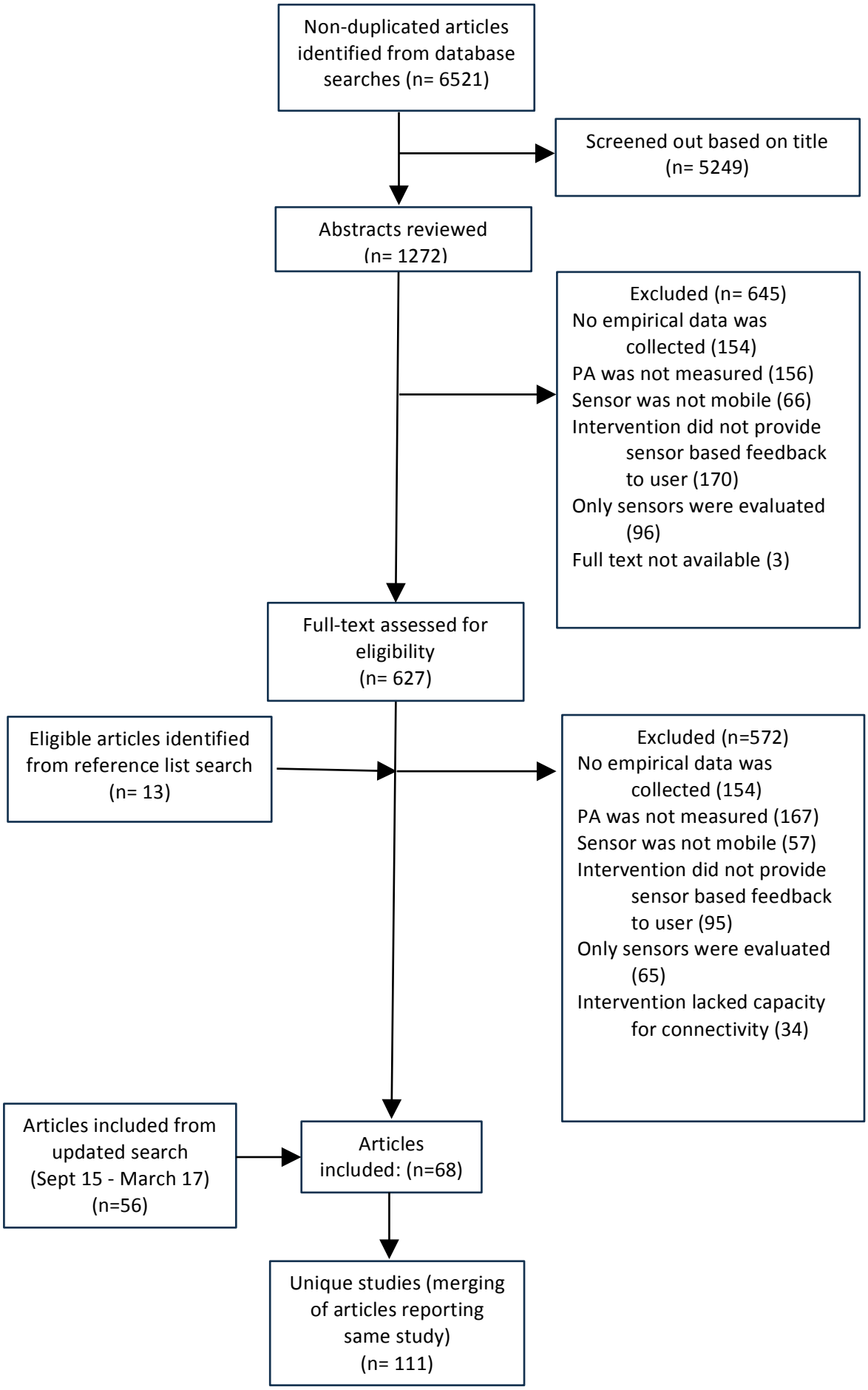


Figure 3: PRISMA flow diagram

Study	Country of Study	N[a]	Age range/mean (SD) [b]	Study Duration [c]
Protocols				
Walters et al 2010	Australia	100	NR	24
Kharrazi et al 2011	USA	60	18-35	4
Pellegrini et al 2012	USA	96	18-60	24
Jimenez Garcia et al 2013	Netherlands	14	NR	4
Geraedts et al 2014	Netherlands	50	70-85	24
Recio-Rodriguez et al 2014	Spain	1553	< 70	52
Clayton et al 2015	Canada	36	50+	11
Cooper et al 2015	UK	488	18-65	12
Granado-Font et al 2015	Spain	70	18+	52
Hurley et al 2015	USA	100	18-60	16
Pellegrini et al 2015	USA	250	18-65	40
Agboola et al 2016	USA	300	18+	26
Amorim et al 2016	Australia	68	18+	52
Duncan et al 2016	Australia	64	18-55	9
Jones et al 2016	USA	200	30-80	13
Ortiz et al 2016	USA	30	13-17	16
Shin et al 2016	South Korea	105	20-39	12
Taylor et al 2016	UK	420	NR	NR
van Nassau et al 2016	International	1000	30-65	52
Brickwood et al 2017	Tasmania	150	60+	52
Ridgers et al 2017	Australia	300	13-14	12
Wolk et al 2017	Germany	120	18-75	4
Completed Trials				
Slootmaker et al 2005, 2009	Netherlands	102	23-39	21
Fujiki et al 2007	USA	4	NR	< 1
Hurling et al 2007	UK	77	30-55	9
Polzien et al 2007	USA	57	41 (8.7)	12
Consolvo et al 2008	USA	30	25-54	3
Faridi et al 2008	USA	30	I: 55(8.7), C:57(10.6)	12
Fujiki et al 2008	USA	E1: 8, E2: 10	E1: 28, E2: 38(7.5)	< 1
Lacroix et al 2008, Goris and Holmes 2008	Netherlands	212	NR	12

Albaina et al 2009	Netherlands	2	65, 73	2
Bickmore et al 2009	USA	8	14-17	9
Fialho et al 2009	Netherlands	12	NR	2
Arsand et al 2010	Norway	12	44-70	12
Mattila et al 2008, 2010	Finland	29	25-54	12
Penados et al. 2010	Netherlands	8	E1: 4-8, E2: NR	1
Lim et al 2011	USA	18	21-53	2
Shuger et al 2011	USA	197	18-64	36
Burns et al 2012	Australia	5	NR	2
Gomes et al 2012	USA	2980	NR	24
Pellegrini et al 2012	USA	51	44 (8.7)	24
Reijonsaari et al 2012	Finland	544	23-64	52
Van Hoye et al 2012	Belgium	227	41 (10.7)	52
Xu, Poole, et al 2012, 2013	USA	1743	10-13	6
Barwais et al 2013	Australia	33	27(4.0)	4
Bentley et al 2013	USA, Sweden	10	NR	8
Chatterjee et al 2012, 2013	USA	E1: 1, E2: 1	E1: 60, E2: 82	8
Fitzsimmons et al 2013	UK	24	68(6)	24
Harries et al 2013, 2016	UK	165	18-40	8
Hirano et al 2013	USA	8	25-70	4
Khalil & Abdallah 2013	Dubai	8	23(2.6)	2
Khan & Lee 2013	Korea	10	NR	10
King et al. 2013, 2016	USA	95	45-81	8
Nakajima et al 2013	Japan	E1: 6, E2: 8	E1: 22-24, E2: NR	3
Tabak et al 2014 a,b	Netherlands	15	66(9.2)	4
Valentin & Howard 2013	USA	6	17-34	< 1
Bond et al 2014, Thomas and Bon 2015	USA	35	21-70	4
Caulfield et al 2014	UK	10	50-80	6
Chen & Pu 2014	Switzerland	36	NR	2
Glynn et al 2013, 2014, Casey et al 2014	UK	90	44(11.5)	8
Miller et al 2014	USA	42	NR	4
Thompson et al 2014	US	49	65-95	52
Thorndike et al 2014	USA	E1: 104, E2: 12	23-37	12
Verwey et al 2014a	Netherlands	20	41-84	12
Walsh et al 2014	USA	74	23-63	4

Zuckerman et al 2014	Israel	E1: 40, E2: 59	E1: 23-54, E2: 20-27	2
Cadmus-Bertram et al 2015a,b	USA	51	I: 59 (6.5), C: 61(7.5)	16
Direito et al 2015	New Zealand	51	14-17	8
Finkelstein et al 2015, 2016	Singapore	800	21-65	52
Frederix, Van Driessche et al. 2015	Belgium	80	I: 63(10), C:58(9)	18
Frederix, Hansen et al 2015	Belgium	140	I: 61(9), C: 61(8)	24
Garde et al 2015	Canada	20	8-13	2
Gouveia et al 2015	International	256	NR	40
Guthrie et al 2015	USA	182	11-14	6
Komninos et al 2015	Greece	20	18-59	NR
Lee, Kim et al. 2015	USA	62	20-68	2
Lee, Cha et al. 2015	Korea	8	22-28	5
Martin et al 2015	USA	48	18-69	5
Munson et al 2015	USA	165	47	12
Rabbi et al 2015	USA	16	NR	14
Verwey et al 2014b, van der Weegen et al. 2015, Verwey et al 2016	Netherlands	199	I: 34(52.3), C: 57(8.3)	36
Wadwha et al 2015	India	30	21-45	8
Wang et al 2015	USA	67	I:49(11.5), C: 47(11.9)	6
Watson et al 2015	UK	65	52(7.4)	52
Broekhuizen 2016	Netherlands	235	60-70	13
Butryn et al 2016	USA	36	40-65	24
Choi et al 2016	USA	30	33.7 (2.6)	12
Ciman et al 2016	Italy	13	24-30	1
Darvall et al 2016	Australia	10	NR	10
Ding et al 2016	USA	19	18-25	4
Fennell et al 2016	USA	15	48.7 (1)	24
Garde et al 2016	Canada	42	9-13	4
Gilson et al 2016	Australia	44	48 (9.8)	20
Glance et al 2016	Australia	353	18-68	16
H-Jennings et al 2016	Singapore	300	18-19	14
Hartman et al 2016	USA	54	60 (5.6)	24
Herrmany et al 2016	International	79	12-72 (47)	12
Melton et al 2016	USA	69	20 (1.7)	8

Patel et al 2016abc	USA	281	40 (11.6)	26
Paul et al 2016	UK	23	56 (10)	6
Quintiliani et al 2016	USA	10	59 (6)	10
Vorrink et al 2016	Netherlands	157	I: 62 (9), C: 63 (8)	52
Walsh et al 2016	UK	58	17-26	5
Yingling et al 2016	USA	8	28-70	2
Ashton et al 2017	Australia	50	18-25	7
Chen et al 2017	Switzerland	36	19-73	8
Chung et al 2017	USA	12	20	8
Gell et al 2017	USA	24	31-78	4
McMahon et al 2017	USA	102	79	24
Neil-Sztramko et al 2017	Canada	19	42 (8.6)	4
Valle et al 2017	USA	35	53 (9.1)	24

Table 2: Study characteristics
a E1, E2 = Separate evaluation trials reported in same study

b Age range if reported, otherwise mean age (standard deviation). I=Intervention group, C= Control group.

NR = Not reported. E1, E2 = Separate evaluation trials reported in same study

c Duration of entire study in weeks, < 1 = less than one week.

5.2.3 Research Designs

Of the 111 included studies (Table 2), 61 studies (55.0%,) used an RCT design. Most of these (40/61, 65.6%) were two-group RCTs; 12 (23.0%) were three-group RCTs and nine (14.8%) were four-group RCTs. Control group participants within RCTs received: standard care or minimal contact/print materials (24/61, 39.3%); active comparison treatments (16/61, 26.2%); non-interactive devices that did not display feedback (11/61, 18.0%); or waitlist/no intervention (10/61, 16.4%). The remaining studies included 23 (22.5%,) repeated measures designs; 11 (9.9%) non-randomised group designs; 10 (9.0%) case studies (six of which included an experimental baseline phase) and four (3.6%) observational studies. Only two studies (1.8%) used rapid research designs: one single case design and one MOST.

- Addition of apps and wearables on non-technology based interventions with health care professionals (Frederix, Hansen et al. 2015, van der Weegen et al 2015, Verwey et al. 2014b, 2016; Watson et al. 2015)
- Addition of gamification features (Zuckerman and Gal-Oz 2014, Direito et al. 2015, Garde et al. 2015, H-Jennings et al. 2016), financial incentives (Finkelstein et al. 2015, 2016, Fennell et al. 2016, Patel et al. 2016a,b,c, Shin et al. 2016) and notifications or short messaging service (SMS) texts (Valentin and Howard 2013) to self-monitoring interventions
- Automation of self-monitoring and goal-setting, including automated activity recognition versus manual input by the user (Bickmore et al. 2009, Duncan et al. 2016), and automated adaptive goal-setting versus standard, static or manual input of goals (Gouveia et al. 2015, Hurley et al. 2015, Lee, Kim et al. 2015, Herrmanny et al. 2016)
- Different social app features that support cooperation or competition (McMahon et al. 2017) or accountability (Chen et al. 2017), social gaming and interaction (Walsh and Golbeck 2014) and personal versus group-based feedback (Harries et al. 2013, 2016, Patel et al. 2016a,b,c)
- Different types of feedback messages, including positive or negative (Nakajima and Lehdonvirta 2013) and novel versus familiar (Gouveia et al. 2015)
- Different prompt frequencies (Thomas and Bond, 2015)

Textbox 2: Intervention components and features investigated for impact on physical activity in included studies

As shown in Textbox 2, studies investigated a variety of intervention components, including the addition of apps or wearables to non-technology-based interventions delivered by health care professionals, and a range of in-app components, such as automated, adaptive goal setting versus static or manual input of goals, and different social components.

5.2.4 Objectives and data collection methods

Table 3 shows the objectives that each study investigated effectiveness, engagement (i.e. device usage and interaction), acceptability (i.e. subjective perceptions and experiences) and/or usability (i.e. characteristics of the device). Almost all studies (96.4%, 107/111) investigated effectiveness, including 14/111 (12.6%) that explored preliminary impact using only descriptive statistics or visual analysis. Only 35/111 studies (31.5%) investigated effectiveness, engagement and acceptability together, and 14 of these (40%, 14/35), did not use inferential statistics analysis to assess effectiveness. Usability was assessed in 16/111 studies (14.4%).

Study	Research Design	Objectives Investigated			
		Effectiveness[a]	Engagement	Acceptability	Usab-ility
Protocols					
Walters et al 2010	2-group RCT	✓			
Kharrazi et al 2011	Non-randomised group design	✓		✓	
Pellegrini et al 2012	3-group RCT	✓	✓		
Jimenez Garcia et al 2013	Repeated measures; randomised crossover design	✓		✓	
Geraedts et al 2014	Repeated measures; pre-post design	✓	✓		
Recio-Rodriguez et al 2014	2-group RCT	✓	✓		
Clayton et al 2015	2-group RCT	✓	✓		
Cooper et al 2015	4-group RCT	✓	✓	✓	
Granado-Font et al 2015	2-group RCT	✓	✓	✓	✓
Hurley et al 2015	4-group RCT	✓			
Pellegrini et al 2015	3-group RCT	✓			
Agboola et al 2016	2-group RCT	✓	✓	✓	
Amorim et al 2016	2-group RCT	✓		✓	
Duncan et al 2016	2-group RCT	✓	✓	✓	✓
Jones et al 2016	2-group RCT	✓	✓	✓	
Ortiz et al 2016	Observational		✓		
Shin et al 2016	3-group RCT	✓	✓	✓	
Taylor et al 2016	Observational		✓	✓	
van Nassau et al 2016	2-group RCT	✓	✓	✓	
Brickwood et al 2017	2-group RCT	✓	✓		
Ridgers et al 2017	2-group RCT	✓	✓	✓	
Wolk et al 2017	2-group RCT	✓	✓		
Completed Trials					
Slootmaker et al 2005, 2009	2-group RCT	✓	✓		
Fujiki et al 2007	Case study with baseline phase	✓ (D)		✓	✓
Hurling et al 2007	2-group RCT	✓	✓		
Polzien et al 2007	2-group RCT	✓	✓		
Consolvo et al 2008	3-group RCT	✓		✓	
Faridi et al 2008	2-group RCT	✓	✓	✓	✓
Fujiki et al 2008	Non-randomised group design	✓ (D)		✓	✓

Lacroix et al 2008, Goris and Holmes 2008	Repeated measures; pre-post design	✓			
Albaina et al 2009	Case study with baseline phase	✓ (D)		✓	✓
Bickmore et al 2009	Repeated measures; randomized crossover design	✓		✓	
Fialho et al 2009	Non-randomised group design	✓		✓	
Arsand et al 2010	Case study	✓	✓		
Mattila et al 2008, 2010	Repeated measures; pre-post design	✓	✓	✓	
Penados et al. 2010	Case study	✓ (D)		✓	
Lim et al 2011	4-group RCT	✓		✓	✓
Shuger et al 2011	4-group RCT	✓			
Burns et al 2012	Case study with baseline phase	✓ (D)	✓	✓	✓
Gomes et al 2012	Repeated measures; pre-post design	✓			
Pellegrini et al 2012	2-group RCT	✓	✓		
Reijonsaari et al 2012	2-group RCT	✓	✓		
Van Hoye et al 2012	4-group RCT	✓			
Xu, Poole, et al 2012, 2013	Repeated measures; longitudinal design	✓	✓	✓	
Barwais et al 2013	2-group RCT	✓			
Bentley et al 2013	Non-randomised group design	✓	✓	✓	
Chatterjee et al 2012, 2013	Case study with baseline phase	✓ (D)		✓	
Fitzsimmons et al 2013	Repeated measures; pre-post design	✓			
Harries et al 2013, 2016	2-group RCT	✓	✓	✓	
Hirano et al 2013	2-group RCT	✓		✓	
Khalil & Abdallah 2013	Non-randomised group design	✓		✓	

Khan & Lee 2013	Case study with baseline phase	✓ (D)		✓	
King et al. 2013, 2016	4-group RCT	✓	✓	✓	
Nakajima et al 2013	Non-randomised group design	✓		✓	
Tabak et al 2014 a,b	2-group RCT	✓	✓		
Valentin & Howard 2013	Repeated measures; cross-over design	✓ (D)			
Bond et al 2014, Thomas and Bond 2015	Repeated measures; randomised crossover design	✓	✓	✓	
Caulfield et al 2014	Repeated measures; pre-post design	✓			
Chen & Pu 2014	3-group RCT	✓	✓	✓	
Glynn et al 2013, 2014, Casey et al 2014	2-group RCT	✓		✓	
Miller et al 2014	Repeated measures; longitudinal design	✓	✓	✓	
Thompson et al 2014	2-group RCT	✓			
Thorndike et al 2014	2-group RCT	✓			
Verwey et al 2014a	Repeated measures; pre-post design	✓	✓	✓	✓
Walsh et al 2014	Repeated measures; randomised crossover design	✓			
Zuckerman et al 2014	3-group RCT	✓	✓		
Cadmus-Bertram et al 2015a, b	2-group RCT	✓	✓	✓	✓
Direito et al 2015	3-group RCT	✓	✓	✓	
Finkelstein et al 2015, 2016	4-group RCT	✓	✓		
Frederix, van Driessche et al. 2015	2-group RCT	✓			
Frederix, Hansen et al. 2015	2-group RCT	✓		✓	✓

Garde et al 2015	Repeated measures; randomised crossover design	✓	✓	✓	
Gouveia et al 2015	Observational	✓	✓		
Guthrie et al 2015	3-group RCT	✓	✓		
Komninos et al 2015	Repeated measures; randomised crossover design	✓		✓	
Lee, Cha et al 2015	4-group RCT	✓		✓	
Lee, Kim et al 2015	Case study; with baseline phase	✓ (D)		✓	
Martin et al 2015	2-group RCT	✓		✓	
Munson et al 2015	3-group RCT	✓		✓	
Rabbi et al 2015	Single Case Design	✓			
Verwey et al 2014b, 2016, van der Weegan et al 2015	2-group RCT	✓	✓	✓	
Wadwha et al 2015	Observational	✓ (D)	✓	✓	
Wang et al 2015	2-group RCT	✓	✓	✓	
Watson et al 2015	2-group RCT	✓	✓	✓	
Broekhuizen 2016	2-group RCT	✓			
Butryn et al 2016	Repeated measures; longitudinal design	✓	✓	✓	
Choi et al 2016	2-group RCT	✓	✓		
Ciman et al 2016	Repeated measures; randomised crossover design	✓ (D)	✓	✓	
Darvall et al 2016	Case study		✓	✓	✓
Ding et al 2016	2-group RCT	✓ (D)		✓	✓
Fennell et al 2016	Repeated measures; crossover design	✓			
Garde et al 2016	Non-randomised group design	✓	✓	✓	✓
Gilson et al 2016	Repeated measures; longitudinal design		✓		
Glance et al 2016	Non-randomised group design	✓			
H-Jennings et al 2016	3-group RCT	✓			
Hartman et al 2016	2-group RCT	✓			
Herrmany et al 2016	3-group RCT	✓		✓	

Melton et al 2016	2-group RCT	✓			
Patel et al 2016abc	4-group RCT	✓			
Paul et al 2016	Non-randomised group design	✓			
Quintiliani et al 2016	Repeated measures; pre-post design	✓ (D)	✓	✓	
Vorrink et al 2016	2-group RCT	✓	✓		
Walsh et al 2016	Repeated measures; randomised crossover design	✓			
Yingling et al 2016	Case study		✓	✓	✓
Ashton et al 2017	2-group RCT	✓	✓	✓	✓
Chen et al 2017	Non-randomised group design	✓		✓	
Chung et al 2017	Non-randomised group design	✓ (D)	✓	✓	
Gell et al 2017	Repeated measures; pre-post design	✓		✓	
McMahon et al 2017	MOST	✓			
Neil-Sztramko et al 2017	Repeated measures; pre-post design	✓	✓	✓	
Valle et al 2017	3-group RCT	✓	✓	✓	
Total		**107**	**58**	**64**	**16**

Table 3: Research designs used in included studies and objectives investigated
[a] D = The study described effectiveness/impact using descriptive statistics or visual analysis, as opposed to inferential statistics.

5.2.4.1 Effectiveness

The majority of studies (101/111, 90.9%) used sensors to measure physical activity. These were most often the in-device sensors used to deliver feedback on physical activity (75/111, 67.6%) (e.g. Fitbit (Caulfield et al. 2014, Chung et al. 2017). Some studies used external sensors (e.g., Acti-Graph GT3X [ActiGraph, Shalimar, FL, USA], Sensewear Armband [BodyMedia, Inc., Pittsburgh, PA], Omron pedometer [Omron Healthcare, Inc., Bannockburn, I]), instead of, or in triangulation with, in-device sensors (26/111, 23.4%). Physical activity data collected via in-device and external sensors included step counts (e.g. Yingling et al. 2016) and activity time (e.g. Van Hoye et al. 2012, Melton et

al. 2016). An external device was significantly more likely to be used in RCTs than in other research designs (X^2 = 7.8, P = 0.005).

Of the 111 included studies, 10 (9.0%) used a questionnaire alone to measure self-reported physical activity, and 17 (15.0%) used a questionnaire to triangulate with sensor data. Questionnaires included the International Physical Activity Questionnaire (IPAQ) (Craig et al. 2003), the Community Health Activities Model Program for Seniors (CHAMPS) (Stewart et al. 2001), the Recent Physical Activity Questionnaire (RPAQ) (Besson et al. 2010), the Godin Leisure-Time Exercise Questionnaire (Godin and Shephard 1985), the Active Australia survey (Australian Institute of Health and Welfare, 2003), the 7-day Sedentary and Light Intensity Physical Activity Log (7-day SLIPA Log) (Barwais et al. 2014), the Yale Physical Activity Scale (YPAS) (Dipietro et al. 1993), and the WHO Global Physical Activity Questionnaire (GPAQ) (Armstrong and Bull, 2006).

5.2.4.2 Engagement

Engagement was measured by 58/111 studies (52.3%) (Table 3), with most (53/58, 91.4%) using device-generated logs to do so. Seven (12.1%) used both logs and self-report questionnaires as a form of triangulation, and five (8.6%) used self-report questionnaires alone. Three dimensions of engagement were identified: frequency or amount of use; depth of engagement (i.e. active vs. passive); and length of use. These are described in Textbox 3.

Frequency or amount of use

- Number of log ins (Reijonsaari et al. 2012, Watson et al. 2015), number of times app opened (Bond et al. 2014, Harries et al. 2013, 2016), number of days device worn (Butryn et al. 2016, Neil-Sztramko et al. 2017, Valle et al. 2017), self-reported frequency of viewing activity trackers (Wang et al. 2015)
- Use of social features, including self-reported frequency of viewing social media messages (Butryn et al. 2016), number of social media messages sent (Hurley et al. 2015, Chen and Pu 2015, Munson et al. 2015, Choi et al. 2016) number of times leaderboard page accessed (Butryn et al. 2016), number of likes/posts on Facebook (Ridgers et al. 2017), number of YouTube video views (Ashton et al. 2017)
- Frequency of use by healthcare professional (Agboola et al. 2016)
- Number of physical activity uploads (Watson et al. 2015)
- Amount of present or missing sensor data (Quintiliani et al. 2016)

Depth of engagement (i.e. active versus passive)

- Whether or not the user manually adjusted pre-set goals (Cadmus-Bertram et al. 2015a,b, Herrmanny et al. 2016), or physical activity levels that were inferred by the device (Duncan et al. 2016)

- Number of missions or challenges completed (Ridgers et al. 2017)
- Logs indicate glancing (5-second intervals with no looking back at step history), review (use or interaction of up to 60 seconds, scrolling through step history), and engagement (use or interaction over 60 seconds, scrolling through step history), and also time between periods of engagement (Gouveia et al. 2015)

Length of use

- Number of times app opened across weeks (Harries et al. 2013, 2016) number of users continuing to post to community board (Butryn et al. 2016), number of days app used post-study (King et al. 2013, 2016)

Textbox 3: Dimensions of engagement assessed by included studies

5.2.4.3 Acceptability

Of the 111 studies included, 64 (57.5%) investigated acceptability (Table 3). Most used questionnaires (41/64, 64.1%), and just over half (34/64, 53.1%) used qualitative interviews or focus groups, either alone or in addition to questionnaires. Questionnaires included a range of standardised established questionnaires (e.g. the IBM Computer Usability Satisfaction Questionnaire (Lewis 1995), the Persuasive Technology Acceptance Model (PTAM) Questionnaire (Connelly 2007), Intrinsic Motivation Inventory (Deci and Ryan 2003), the Fun Toolkit (Read 2008) and the Working Alliance Inventory (WAI) (Horvath and Greenberg, 1989), or questionnaires developed especially for the study (e.g. Bickmore et al. 2009, Bentley et al. 2013). A few studies employed user logs (7/64, 10.9%): four used device-generated usage logs as a "proxy" (Wadhwa et al. 2015) of users' interest (Gouveia et al. 2015) or preferences (Ding et al. 2016); three used user-entered text (e.g. the content of social media messages to understand the types of social support that users experienced (Poole et al. 2013, Chen and Pu 2015, Munson et al. 2015), and digital diary entries to understand experiences of using the device (Chen and Pu 2015). Studies that used text-based logs also employed face-to-face qualitative methods (i.e. interviews, focus groups) or questionnaires, in addition to collecting log data. Five dimensions were identified in relation to measuring acceptability: appreciation; perceived effectiveness and usefulness; satisfaction; intention to continue use, and social acceptability. These are described in Textbox 4.

5.2.4.4 Usability

Usability was investigated by 16 studies (14.4%): nine (56.3%) used questionnaires (e.g. the System Usability Scale (Brooke)); four (25.0%) used

interviews; two (12.5%) used focus groups; and one (6.3%) (Fujiki et al. 2008) used observation of participants completing timed tasks. Three dimensions were identified in relation to assessing usability: burden of device wear and use, interface complexity, and perceived technical performance. These are described in Textbox 5.

Appreciation

- Appreciation or liking of the app (Albaina et al. 2009, Komninos et al. 2015, Ciman et al. 2016)
- Whether the app or wearable was perceived as enjoyable, fun, entertaining (Fialho et al. 2009, Garde et al. 2015, Lee, Kim et al. 2015, Ridgers et al. 2017)
- Whether app or wearable was pleasant (Bond et al. 2014), attractive or visually appealing (Ashton et al. 2017)
- What was "missed" about feature once withdrawn (Lee, Cha et al. 2015)
- How user 'felt' about the app or wearable and its components (Consolvo et al. 2008, Albaina et al. 2009, Lim et al. 2011, Nakajima and Lehdonvirta 2013)
- Users' interest and preferences (Gouveia et al. 2015, Wadhwa et al, 2015, Ding et al. 2016)
- Teachers' perceptions of whether the app or wearable appealed to students (Ridgers et al. 2017)
- Self-reported motivation to pay attention (Lee, Kim et al. 2015)
- Perceived advantages and disadvantages of using the app or wearable (Zuckerman and Gal-Oz 2014, Amorim et al. 2016)

Perceived effectiveness and usefulness

- Users' views on whether the app or wearable increased, or will continue to increase and promote, physical activity (Fujiki et al. 2008, Albaina et al. 2009, Arsand et al. 2010, Lim et al. 2011, Hirano et al. 2013, Bond et al. 2014, Verwey, et al. 2014a , Frederix, Hansen et al 2015, Garde et al. 2015, 2016, Komninos et al. 2015, Agboola et al. 2016, Ding et al. 2016)
- Practice nurses' perceptions of effectiveness for patients (Verwey et al. 2014b)
- Users' perceived usefulness or helpfulness of the app or wearable (Albaina et al. 2009, Fialho et al. 2009, Bond et al. 2014 , Cadmus-Bertram et al. 2015a,b) and its components (Wang et al. 2015, Agboola et al. 2016, van Nassau et al. 2016, Neil-Sztramko et al. 2017, Valle et al. 2017) in self-monitoring (Duncan et al. 2016), supporting fitness and physical activity (Direito et al. 2015, Wang et al. 2015), and supporting them to stay motivated (Gell et al. 2017)
- Users' perceived persuasiveness or helpfulness of the app or wearable in achieving goals
  (Ashton et al. 2017)
- Ability of the app or wearable to provide answers to health-related questions (Ashton et al. 2017) and insight into physical activity or health conditions (Agboola et al. 2016)
- Health care professionals' perceptions of the usefulness of information about patients' physical activity or health condition and whether it supported engagement with patients' home care (Agboola et al. 2016)

Satisfaction

- General user satisfaction (Arsand et al. 2010, Walters et al. 2010)
- User satisfaction with number of reminder short messaging service or calls received (Wang et al. 2015)
- User satisfaction with length of intervention (Ashton et al. 2017, Ridgers et al. 2017)
- User satisfaction with level of personalization (Lee, Kim et al. 2015) and feedback provided by the app or wearable (Duncan et al. 2016)
- Likelihood of users recommending the app or wearable to a friend or other people (Cadmus-Bertram et al 2015a,b, Butryn et al. 2016, Chung et al. 2017, Neil-Sztramko et al. 2017, Valle et al. 2017)
- Satisfaction with different components or features (Cadmus-Bertram et al 2015a,b, Frederix, Hansen et al. 2015, et al. 2015, Garde et al. 2016, Gell et al. 2017, Neil-Sztramko et al. 2017, Valle et al. 2017)
- Likelihood of physicians recommending the app or wearable to patients (Jones et al. 2016)

Intention to continue use:

- Intention or willingness to use after the study (Albaina et al. 2009, Harries et al. 2013, 2016, King et al. 2013, 2016, Bond et al. 2014)

Intention to continue use (cont'd).

- Intention to continue use if user had to pay for the app or wearable (Quintiliani et al. 2016) or intention to purchase the app after the study (Butryn et al. 2016, Ashton et al. 2017).
- How regularly the user intended to use the app or wearable after the study (Duncan et al. 2016, Jones et al. 2016)

Social acceptability

- Whether the app or wearable was noticed and remarked upon by others (Lim et al. 2011), or prompted discussed with others (Agboola et al. 2016)
- Whether the app or wearable was used by important others (Albaina et al. 2009.)
- Users' attitudes towards sharing data with others (Munson et al. 2015)
- Social encouragement (Garde et al. 2015) and social support received when using (including via) the app or wearable (Xu et al 2012, Poole et al. 2013, Miller and Mynatt 2014)
- Level of social bonding between the user and virtual coach (Bickmore et al. 2009),
- Users' preferences between in using individual versus social features (Chen et al. 2017)
- Whether context-aware notifications were received at a socially acceptable time and place (Glance et al.) or interfered with users' daily activities (Lee, Cha et al. 2015)

Textbox 4: Dimensions of acceptability assessed in included studies

Burden of device wear and use

- Ease of wear (Garde et al. 2016), burden or restriction in wearing device, physical discomfort (Darvall et al. 2016, Yingling et al. 2016), usability regarding the device size (Burns et al. 2012), suggestions for alternative wear locations (Cadmus-Bertram et al. 2015a,b)
- Ease of use (Albaina et al. 2009, Granado-Font et al. 2015, Ding et al. 2016) when syncing to web-based databases (Darvall et al. 2016, Yingling et al. 2016) and when charging the device (Burns et al. 2012)
- Whether device interfered with daily activities (Frederix, Hansen et al. 2015)

Interface complexity

- Complexity and intuitiveness (Fujiki et al. 2007), accessibility (Yingling et al. 2016), and comprehension of physical activity feedback (Ashton et al. 2017)
- Ease of reading information (Frederix, Hansen et al. 2015)
- Difficulties using the interactive interface, users' speed when completing in-app tasks (Fujiki et al. 2008)

Perceived technical performance

- Users' perceptions of the accuracy of the app or wearable in recognizing or inferring their physical activity (Fujiki et al. 2007, Darvall et al. 2016, Duncan et al. 2016).
- Technical difficulties or barriers encountered by users (Verwey et al. 2014a, Cadmus-Bertram et al. 2015a,b)

Textbox 5: Dimensions of usability assessed in included studies

5.3 Discussion

Of 111 studies included, around half were published between 2015 and 2017, 55% were RCTs, and only two studies used rapid designs. Almost all studies measured physical activity objectively using sensors (either in-device or external), with RCTs more likely to employ external sensors (accelerometers). Less than a third of studies investigated effectiveness, engagement and acceptability together. Studies that measured engagement mostly used device-generated logs, while studies exploring acceptability most often used questionnaires and/or qualitative methods. Dimensions of engagement and acceptability were tentatively proposed, including the frequency, depth, and length of engagement, and, in relation to acceptability: appreciation; perceived effectiveness and usefulness; satisfaction; intention to continue use; and social acceptability. A small number of studies explored usability of the device (including burden of sensor wear and use, interface complexity, perceived technical performance) using questionnaires, qualitative methods, or participant observation.

The fact that more than half of the included studies were published between 2015 and 2017 demonstrates that research into the impact of physical activity apps and wearables is a growing area of interest, underscoring the timeliness of this scoping review. Despite this, only two studies were found to have used rapid research designs that have been recommended for evaluating mobile health technologies (single case design (Rabbi et al. 2015), and the MOST approach (McMahon et al. 2017). A low uptake of rapid research designs was similarly reported in a recent review of clinical mHealth app evaluations (Pham et al. 2016); however, while the vast majority of evaluations of clinical apps were RCTs, findings from the current review indicate that evaluations of physical activity apps and wearables use alternative research designs (including repeated measures designs, non-randomised group designs, case studies and observational studies) more often. This may reflect the interdisciplinary nature of the review, and the view held by some HCI researchers that RCTs, as well as being impractical and resource intensive, are of limited usefulness (Klasnja et al. 2011).

It is surprising that few studies used single case designs and new factorial approaches, as it has been suggested that mHealth technologies can support the data collection procedures and experimental set-up these research designs require (i.e. frequent measurement and facilitate several experimental conditions such as app versions) (Dallery et al. 2013, Hekler et al. 2013, Michie et al. 2016). Rabbi et al. (2015) did take advantage of technological features to support their SCD evaluation of a physical activity app; this study used frequent data collection from internal smartphone sensors to support the requirements of the SCD design. In their study of a wearable using MOST, however, McMahon et al (2017) varied aspects of the intervention that was delivered by healthcare professionals, as opposed to demonstrating the effectiveness of different device components. Thus, studies could be doing more to capitalize on technology features to support and automate their research designs.

Many evaluations of physical activity apps and wearables appear to be taking advantage of efficient data collection methods: two-thirds of studies employed in-device sensors in smartphones and wearables to measure physical activity. The fact that RCTs used external, validated sensors more often than other study designs exacerbates their inefficiency (e.g. through adding extra resource costs, Ferguson et al. 2015). Furthermore, using external sensors often involves measurement procedures that may reduce the generalisability of findings to real world contexts (e.g. requiring participants to wear additional devices and visit the lab). The coupling of gold standard RCTs, and sensors with established validity, indicates a well-founded concern for methodological rigour. However, how to balance this need for rigour with efficiency, requires further exploration.

In addition to effectiveness, assessing user engagement and acceptability are important to: generate a better understanding of overall impact; explain variation in outcomes; and reveal (potentially interactive) influences on effectiveness (Oakley et al. 2006, Grant et al. 2013). Despite this, only around a third of studies (32%) investigated all three objectives together. Furthermore, 40% of these did not use inferential statistics to assess effectiveness (instead using descriptive statistics and visual analysis), and almost 20% of all studies contained fewer than 13 participants. These preliminary, small-N studies are typical of iterative HCI research focussed on developing novel technologies (Kay

et al. 2016), yet they are unlikely to be sufficiently powered to test important hypotheses on mediators of effectiveness (Petticrew et al. 2011, Moore et al. 2015). Although the current study did not explore which statistical analyses were undertaken, Bayesian methods are considered a promising approach for mHealth evaluations (Dobkin and Dorsch 2011, Hekler et al. 2016, Michie et al. 2017), and can be used to investigate mediating variables in small-N studies (Miočević et al. 2017). As such, Bayesian methods could be key when exploring results from early developmental evaluations to reveal potential relationships between mHealth engagement, acceptability and effectiveness.

A few studies assessed usability. In line with other conceptualisations of usability (i.e. whether the device or app is easily used to achieve specified goals successfully and quickly (ISO 1998, Quesenbery 2003), usability was distinguished from acceptability by considering it to be a characteristic of the device. Understanding the degree to which usability varies across users and interacts with context to ultimately influence effectiveness (as opposed to being a stable device characteristic) will dictate whether it should be assessed during effectiveness evaluations, alongside acceptability and engagement. It may be that usability is more important to explore and optimise during the development phases of physical activity apps and wearables.

The screening process in this interdisciplinary review involved a very high number of abstracts and full papers being read to identify the final studies for inclusion. Many of the articles retrieved from the database searches had ambiguous titles; and many authors omitted key study details from their abstracts. Furthermore, data extraction from the full text articles involved negotiating different publication formats across disciplines. These challenges meant the review process was far more time-consuming than originally envisaged. Currently, HCI studies are not required to follow heath science reporting guidelines that promote the inclusion of specific study details in titles and abstracts (e.g. Schulz et al. 2010). Standardized reporting drawing on existing guidelines (e.g. CONSORT-EHEALTH, Eysenbach et al 2011) would allow different disciplines to more easily synthesise the large amount of research that is being conducted in this area, and would also aid current efforts to develop

automated processes to increase the accessibility of evidence from digital health publications (e.g. Michie et al. 2017).

5.4 Conclusions

Despite the rapid increase of evaluations of the impact of physical activity apps and wearables, few are optimised in relation to efficiency and assessment of the key constructs of engagement and acceptability, as well as effectiveness. Health and HCI researchers need to be supported in making greater use of rapid research designs (e.g. single case designs), in-device sensors and user-logs to collect effectiveness, engagement and acceptability data. The difficulties encountered in conducting this interdisciplinary review also highlight the need for standardized reporting guidelines. These would facilitate the synthesis of evidence across health and HCI disciplines, and thus support rapid advancement of our understanding of the extent to which apps and wearables can support users to become more physically active.

Future research should investigate why recommended rapid research designs are not yet being frequently adopted. Qualitative explorations of researchers' perceptions, daily research practices and experiences would allow understanding of the practical challenges in using rapid designs, which may differ across disciplines. This formed the basis for Study 2, which is presented in the next chapter. Furthermore, given the need for efficiency in mHealth evaluations, studies should explore the extent to which rapid designs can be supported by mHealth technologies and automated (Riley et al. 2013). The rapid research designs found in the review were single case design (SCD) and MOST approaches, indicating that these are feasible for physical activity apps. It is also encouraging that despite not following the principles of SCDs, a number of studies did resemble SCDs (i.e. case studies with baseline phases). Automated SCD studies are discussed in Study 3, which is reported in chapter 8 and 9.

Chapter 6 Interdisciplinary perspectives on scoping review findings

6.1 Introduction

Fifteen interviews were conducted to understand current practices of academic researchers and industry professionals and how they relate to the findings of the scoping review. This chapter presents the perceptions and experiences of health and HCI researchers, industry data scientists and other industry professionals in relation to the following issues raised by the scoping review:

1. The majority of studies used randomised controlled trials (RCTs)

2. Most studies used in-device sensors, with RCTs more likely to use external, research-grade sensors

3. Most studies employed user logs to assess engagement

4. Few studies used logs to assess acceptability

This chapter also outlines further research designs and methods that researchers and industry professionals used beyond those identified in the scoping review.

6.2 Participants

The characteristics of participants (who were anonymised using pseudonyms) and where they worked are described in Table 4. Interviewees were mostly male (12/15). Eight were from academia, seven were industry professionals However, it proved hard to categorise academic researchers as specialists in specific disciplines (i.e. a 'health' researcher or 'HCI' researcher') as not all exclusively identified with a single discipline. This appeared to be particularly problematic for those involved in HCI research; when one researcher was asked which discipline she felt she belonged to, she indicated that it was complex as she had moved between areas of research and that she "hate[s] that question" [Liz, HCI Researcher]. Another researcher described himself as a "hybrid" between HCI and psychology, and recounted that he often wore different "hats" (i.e. altered

his focus and perspective) when critiquing different app designs, or the validity of studies [Michael, HCI Researcher].

Several participants (8/15) reported involvement in the design and/or evaluation of physical activity apps and wearables. The other behaviours targeted included nutrition and diet (4), alcohol (2), medication adherence and medical technology use (2), and smoking (1). Other projects included the apps for general habit formation (2) and emotional regulation (2).

Five participants explicitly reported that they did not evaluate the effectiveness of apps, of these three were industry professionals and two were HCI researchers. Industry professionals who had undertaken effectiveness evaluations included one who provided app design and evaluation as a service to other industry professionals [Mark], and one whose company was embedded within a research lab.

Name	**Gender**	**Sector***	**Discipline/ Profession**	**Areas of Interest**	**Effectiveness evaluation?**
George	M	Industry	Product Designer	Alcohol reduction, physical activity	N
Michael	M	Academia	HCI Researcher (Senior)	Nutrition	Y
James	M	Industry	CEO	Physical activity	N
Joseph	M	Academia	Health Researcher (Early Career)	Physical activity	Y
Maria	F	Academia	Health Researcher (Early Career)	Alcohol reduction	Y
Tom	M	Academia / Government	Health Researcher (Senior)	Nutrition, emotional regulation	Y
Mark	M	Industry	Product Designer/ Academic (Senior)	Physical activity, range of behaviours	Y
Pat	M	Industry	CEO	Nutrition (general habit formation)	Y
Gillian	F	Academia	HCI (Senior)	Physical activity	Y
Aaron	M	Industry	Industry Data Scientist	Physical activity, mental health, smoking	Y
Steve	M	Industry	CEO	Nutrition	N

				(general habit formation)	
Chris	M	Academia	Health Researcher (Senior)	Physical activity	Y
Liz	F	Academia	HCI Researcher (Senior)	Physical activity, medical technologies	N
Fred	M	Academia	HCI Researcher (Senior)	Physical activity	N
Daniel	M	Industry	Industry Data Scientist	Physical activity	Y

Table 4: Characteristics of participants in Study 2
*current employment

6.2.1 Use of randomised controlled trials

The scoping review found that most studies were RCTs; only two employed rapid research designs. Some researchers in health and HCI had conducted RCTs, as had some industry data scientists, however product designers and CEOs had not. One health researcher [Maria Health Researcher] had used the MOST design. Two themes emerged in relation to the scoping review finding that the majority of studies continue to use RCTs: RCTs as the 'gold standard' evaluation design, and suitability for evaluation of apps and wearables.

6.2.1.1 Gold standard evaluation design

HCI researchers, health researchers and industry data scientists described RCTs as "the gold standard" for demonstrating whether or not an intervention is effective. Participants spoke about the "validity" of RCTs and how they were an "established" means of achieving valid results [Aaron Industry data scientist]. Participants across different disciplines also recognised RCTs as "traditional" [Daniel industry Data Scientist] and "classical" [Michael HCI Researcher], particularly when they consisted of only two experimental conditions [Tom Health Researcher, Fred HCI Researcher] or assessed an intervention as a whole "package", as opposed to investigating individual intervention components [Maria Health Researcher].

Participants compared the use and nature of RCTs in "pharmaceutical science" and digital health [Joseph, Health Researcher]. One HCI researcher felt that models of evaluation for any kind of digital health technologies were "driven by

pharma" and based on the RCTs used in drug trials [Liz, HCI Researcher]. Interviewees described instances where behaviour change apps may ultimately be used by healthcare organisations that only accept evidence of effectiveness from RCTs. One industry data scientist, Aaron, provided the example of apps being prescribed by healthcare professionals, and suggested that "regulation" in these settings enforced and maintained the requirement for RCTs, as they were seen as the only acceptable means of demonstrating validity and rigour:

> I had discussed RCTs with different people, both people in hospitals as well as people in pharmaceutical domains. Both of them struck me as places that are highly reluctant to deviate from an established method, particularly because of regulation. So both... they said that unless they run RCTs they won't get anything through the door because of how highly regulated they are... so imagine getting a digital behavioural intervention to be prescribable... The answer was always that RCTs are the golden standard and that, due to regulatory requirements, they won't be able to demonstrate the validity of approaches unless they follow that [Aaron, Industry Data Scientist]

Aaron had discussed the requirement of RCTs with healthcare professionals in hospitals; health insurance companies were also perceived as requiring RCTs to demonstrate the effectiveness of behaviour change technologies. Michael, an HCI researcher, stated that using RCTs was the only way health insurance companies would support the use of these technologies:

> For the health insurances to actually pay for a product...basically what you need is controlled trials, that is the only way that you can, that you can get the health insurance in [this country] which is private, how you can get them to pay [Michael, HCI Researcher]

Thus, interviews felt both public (e.g. hospitals) and private (e.g. pharmaceutical and health insurance) organisations continue to regard RCTs as the gold standard and will neither adopt not financially support technologies that have not been assessed using an RCT design.

6.2.1.2 The suitability of RCTs for evaluating behaviour change apps and wearables

Industry product designers and CEOs did not generally discuss the suitability of RCTs[12]. Among health researchers, HCI researchers and industry data scientists,

[12] This refers to product designers and CEOs: data scientists *did* discuss the suitability of RCTs.

however, there were mixed views on the suitability of RCTs for evaluating behaviour change apps and wearables. One health researcher, Chris, felt that there was a "need" for RCTs because of the "control" they provide [Chris Health Researcher]. Similarly, Michael [An HCI Researcher] had used a lab-based RCT to assess a behaviour change technology targeting nutrition because it had been important to "keep as much constant as we possibly could" between participants. This involved setting up a lab "so it looked like a restaurant... as much as possible as a reasonably natural eating environment" to enhance the validity of conclusions from the experiment.

Other health and HCI researchers felt that while it "makes sense" [Joseph, Health Researcher] or they could "see the argument" for the use of RCTs in pharmaceutical domains [Liz, HCI Researcher], these were not always possible or appropriate for behaviour change apps and wearables. Joseph spoke about the impracticality of controlling participants' use of other apps:

> we can't really enforce and control whether some people have used other apps... You know, so co-interventions, we can't control co-interventions, whether they decided to try, or download other apps, in addition to the one that we gave them [Joseph, Health Researcher].

Another issue reported when conducting RCTs was the evolving nature of technology, as RCTs typically require an intervention to be kept constant throughout a trial (Chorpita et al. 2005). Participants described problems arising from continuous updates to the operating system (i.e. basic software or platform) on which their app was based. Aaron recalled discussions with healthcare organisations that required RCTs, where he had highlighted that a changing app may reduce validity:

> I had many discussions where I emphasised even the smaller details, like the fact that the timespan required for an RCT is incompatible with stopping developing your software because what if, in the time that you're running your RCT, Google releases a new version of android? You need to keep working on your software, and then how do you marry the fact that, during your RCT, you are changing your software and does that invalidate your results if you were to start over? [Aaron, Industry Data Scientist].

One HCI researcher, Fred, described his experience of an operating system update that required the research team to update an installation manual for a Fitbit app. Although in this instance the device itself did not require to be updated, Fred suggested that device updates were likely to occur in future and that different study designs beyond RCTs were needed to accommodate this:

> ...don't change the technology after the beginning of the study. You know, that's standard... We had this in another study... what happens when the technology is changing? We... have them install Fitbit app on their smartphone or on their Windows machine, and then more and more people use Windows 10 and our instructions and guidelines and the manuals for how to install the app didn't work anymore and this is a very classic example.... It was trivial. In this case it was sufficient to just update the manuals. What would happen when the device that we gave them is no more supported? So, it's really... this can happen and this will happen and I don't have a solution for that. I mean in the end you really need different study designs. [Fred, HCI Researcher]

Interestingly however, the evolving nature of technologies did not only pose problems for those conducting standard RCTs. Maria, a health researcher, recounted how changes to Apple's operating system had presented challenges when conducting a MOST factorial trial. As MOST requires the development of several different app versions to constitute different experimental conditions, all versions had to be checked to ensure their compatibility with the operating system:

> you know, a new version of IOS comes out and then you need to run all those checks again... And there were, you know, things that weren't problems before and then a version changed and if you tried to enter something or other there was a bug or it didn't work [Maria, Health Researcher].

Maria highlighted that these changes to operating systems are "outside" of researchers' "control", and Fred [HCI Researcher] similarly described how the nature of the updates, and whether they were likely to occur, were difficult to anticipate as "you could not know before" conducting the study.

HCI researchers also expressed views about the suitability of RCTs for typical HCI studies. One noted that RCTs were not amenable to the exploratory nature of the research questions often addressed by HCI:

> you've got to have a very well defined question, before that kind of clinical trial is an appropriate method... So if your question is more exploratory... you know, how could this design be improved... in terms of basic usability, in terms of acceptability, in terms of the way it might fit into people's lives... how you personalise it for people with different motivations, or different interests... RCTs don't get those questions at all [Liz, HCI Researcher].

In contrast, another HCI researcher [Gillian] felt that RCTs were "hugely beneficial" for the HCI community, because the large sample sizes available when conducting pragmatic app store ("in the wild") RCTs enabled HCI researchers to advance theories on social norms and how to support 'nudging' of behaviours. Gillian also felt that pragmatic RCTs provided further opportunities to explore how devices are used outside of HCI labs. Pragmatic RCTs are further discussed in section 6.3.1.

6.2.2 In-device vs external sensors

Given the potential for in-device sensors to measure physical activity outcomes, it was interesting to find in the scoping review that some researchers still used external research-grade accelerometers, particularly within RCTs. In the interviews, the relative *accuracy* of external sensors emerged as an overarching theme that appeared to be interrelated with different constructs, reliability, user burden and engagement, and data availability. Specifically, participants appeared to feel there were trade-offs between gold standard measurement accuracy and these constructs.

6.2.2.1 Measurement accuracy

Participants felt that there were differences in measurement accuracy between research-grade and in-device sensors. Research-grade sensors were recognised as more 'precise' and as the 'gold standard' in measuring physical activity by both health researchers [Joseph and Tom] and HCI researchers [Fred, Liz], who felt that they produced 'validated data'.

Reflecting on the finding that many studies still employed external accelerometers, health researcher interviewees in particular felt that the considerations of the likelihood of publication of their research in high quality journals may have influenced researchers' decisions over which sensors to use:

> there might be a bit of... technocracy in terms of journals and reviewers, and what happens if some other researcher in the physical activity space might expect to see.... that the physical activity data is being measured by a research grade, by the gold standard, let's call it that. As opposed to just relying on smartphone inbuilt sensors, or trackers [Joseph, Health Researcher]

> you're still always thinking about what will be accepted in peer review to get this article published so the tendency is to, and like in the case of Actigraphs... We know it's research grade. We know its validity. We know what it's capable of. With more strengths than weaknesses everybody kind of accepts it as sort of the gold standard for physical activity outcome [Tom, Health Researcher]

The scoping review found that studies that employed RCTs were more likely to employ research-grade sensors. Another health researcher (who had not employed sensors herself) suggested that researchers might plan their methods based on perceived journal acceptance and risk of not being published:

> If you are going to the additional effort of a full RCT you want to make sure that your measures hold up to any criticism... you don't want a review journal to come back and say oh well, you know, I don't trust your measurements, and not to have that as something that could factor in or be a limitation in your study [Maria, Health Researcher]

Participants were clear that research-grade devices were expensive. However, health researchers tended to value their use "whenever we have the money" [Chris, Health Researcher], whereas some HCI researchers felt their use was not warranted in the type of work they do:

> I don't do the kind of work that would justify investing in research grade sensors, because I don't use RCTs [Liz, HCI Researcher]

6.2.2.2 Reliability

A further factor interrelated with accuracy was measurement reliability, or "consistency" [Gillian, HCI]. Both health and HCI researchers described how the reliability of in-device sensors was perhaps more important than their accuracy:

> it seems like people, researchers... like there's more acceptance towards the fact that, okay, if we use these commercially available smart watches, or sensors, or rely on the phone...it won't be as accurate as an accelerometer. But it should be reliable enough to at least detect changes, which are what we most, most research

> questions are trying to answer, right... maybe we just need to know if things change...in terms of reliability [Joseph, Health Researcher]

> Even if the absolute value may be wrong…it's wrong but still you see deviations from day to day. So, the intrapersonal reliability is, indeed, quite high so you can see changes, and that is really my key point [Fred, HCI Researcher].

6.2.2.3 User burden and engagement

A further factor interrelated with accuracy was user-burden. Participants perceived there to be differences between devices in their degree of user burden and "hassle" [Daniel, Industry Data Scientist]. It was felt that although accurate, research-grade devices were more burdensome than in-device sensors and that participants might stop using them. Health researchers noted that using research-grade devices meant "you end up with people wearing two devices" which might be "weird" for the participant [Tom, Health Researcher]. Maria felt that participants may "be more likely to engage" in an intervention if they did not have to wear this "additional" sensor and that it was important not to "lose the user" from the study altogether [Maria, Health Researcher].

HCI researchers similarly suggested that research-grade devices were "obtrusive" and participants would not wear them long term [Fred, HCI Researcher]. Fred advocated the use of commercially available devices such as Fitbit, however another HCI researcher noted that even these could be intrusive. Liz recounted her own experience of finding Fitbit notifications "disruptive" and "distracting" during meetings which lead to her abandon the device [Liz, HCI Researcher]. Liz explained user burden and engagement therefore depend on the context in which a device is used (i.e. as opposed to simply being properties of the device). Fred, on the other hand, felt that engagement, and consequently the accuracy and reliability of measurement data, is governed by whether users feel a measurement device is actually "appropriate" for their own use.

6.2.2.4 Data availability

There was a desire amongst both researchers and industry professionals to explore the use of in-device sensor data because it was already being collected. Mark, an industry professional who evaluated effectiveness stated that his company "want to use... to look at that data" from in-device sensors "because

we're collecting that data" *(*i.e. for intervention purposes). Fred, an HCI researcher, felt that the "availability" of long-term data from commercial devices was more important and useful than short-term accurate data from research-grade sensors.

Interestingly, health researchers highlighted that the availability of in-device sensor data could be limited, because industry professionals intentionally restricted access to information on their commercial products. One health researcher [Joseph] recounted how his research team "couldn't extract" usage logs or sensor data because iPhones were "closed systems". This meant he was not able to view or change the apps' code and logs or determine specifically what data was to be collected. Another health researcher felt beyond outcome data itself, industry professionals did not make the algorithms that were used to calculate outcomes such as physical activity, available. This meant their accuracy and validity was unknown:

> Fitbits and everything else... we importantly don't know the underlying algorithm by which they [industry professionals] compute calorie expenditure and that sort of thing... [Tom, Health Researcher]

6.2.3 Assessing engagement through device generated user logs

The scoping review found that the majority of studies that assessed users' engagement with physical activity apps and wearables employed user logs to do so, as opposed to self-report methods. Within the interviews, two related themes emerged: log data increases objectivity; and logging software.

6.2.3.1 Log data increases objectivity

Industry professionals and health researchers reported advantages of log data in providing a more objective measure of engagement than self-report methods. One industry professional, Mark, felt that participants may report very different usage patterns to those found within log data. He gave the example of users stating they "use it all the time... and then you look at the data" to find that they do not. Similarly, a health researcher, Tom, suggested logs can more accurately capture the details of the context (how and what) of engagement than self-report:

> You certainly prefer the actual log data to the self-report of how they used the device. I mean, you have, why rely on retrospective self-report when you have prospective objective data on exactly how it was used? So whenever that's available I think that's always the preference to be able to pull when did people use it, how did they use it, what did they do on it; all the data that you can collect from that. [Tom, Health Researcher]

6.2.3.2 Logging software

Participants described how they collected log data using logging software. Participants either used "internal tooling" developed by their own company [Aaron, Industry Data Scientist] or freely available software such as Google Analytics. Both industry professionals and academic researchers felt that freely available software made engagement data easy to collect:

> there are third party tracking tools that make this stuff, and first party, actually, from Apple and Google, they make this stuff really, really easy [James, CEO]

> Well, we've always kind of worked with Google Analytics, because these tools are out there... they're so comprehensive that's it's like, and so easy to implement, there's no point in not using it. And it's free as well... I still come across a lot of papers... they haven't even got usage data and it always frustrates me knowing that Google have it for free and it's quite easy to have your website tracked by Google analytics [Chris, Health Researcher]

Although interviewees described logging data as easy to use, health researchers in particular recounted how they had to rely on colleagues (often from other disciplines) to help them implement the software and prepare the data for analysis. Chris indicated that a developer or 'programmer' set up the logging software for their intervention and that he himself did not know how to use it:

> I mean, I don't know how to do it right, but I've got a programmer and so I presume everybody else will have a programmer too potentially, so, from that point of view I wouldn't think it's very hard, but don't ask me how to do it. [Chris, Health Researcher]

Another health researcher, Joseph, had not used engagement logs before and felt he would need help from those in other disciplines, specifically HCI, to do this. Joseph went on to say that health behaviour change researchers do not

necessarily consult those in other disciplines despite them having knowledge of how to use logging tools:

> I'm not familiar with a lot of... there's limited knowledge in terms of the instruments that are already there... we don't, with behavioural science, I think, like, we're not using some other expertise on this, when we assess these things [Joseph, Health Researcher]

Another health researcher recounted that a colleague had helped to clean and prepare log data for analysis. She felt her colleague was familiar and skilled with the kind of processes required (such as to "write" and "check" necessary code), which allowed him to clean the data more quickly than she felt she could have:

> I've got to admit I worked on this with [another researcher] at the time and he did the kind of clean-up of the app usage data, but I remember he basically, he set a macro in Excel to run and then left it for like a day and a half.... So like once he has run the macro it was then fairly straightforward to run... but it did take him a while to write that and to check it.... And he is also quite, you know, he is good at things like that, I think it would have taken me particularly longer. [Maria, Health Researcher]

Overall, health researchers felt that working with colleagues with technical skills (either developers or those in HCI researchers) was needed both to collect data, and prepare it for analysis. Health researchers further explained that it was not always easy to analyse and interpret the data collected by logging software. Chris explained that logging software generated a "massive" volume of data and that although it was valuable, it required effort to use (it is "gold and it's free but it's notoriously hard work" to "dig out" the information needed). Chris also noted that the data was "very unstructured...not like a questionnaire". The overwhelming amount of data, its unstructured nature, and its overall "complexity" were felt to make it difficult to pull out specific data of interest (i.e. to "query the database"). Chris described the need to understand and "figure out" how to do this:

> ... how should we even query the database. How should we bring this back to us, and how should we make sense of it... there's a lot of interventions we've got that go on for three months or so and then there's a six month follow up, and we have weekly data. So how do we get that to show what actually happens during the intervention phase, what happens after the intervention phase [Chris, Health Researcher].

Chris noted that part of the problem was because "you can ask so much" and "there's so many different ways" to interpret the data [Chris, Health Researcher]. Another health researcher, Maria, expressed the need for more guidance on what the "best" way was, and what specifically should be measured in relation to engagement:

> ... think sort of almost a clearer, like, toolkit on what we needed to measure..... we used log data and it was very much, we were very unclear what the best way of evaluating those things were. [Maria, Health Researcher]

Nevertheless, researchers and industry professionals reported using innovative features of logging software that helped interpret user log data, such as visualisations of engagement and attrition [Pat, CEO], which were "brutal" in showing the reality of declining engagement [Chris, Health Researcher]. Pat described how some Google tools can not only be used to assess "churn" (i.e. when users stop engaging with the app), but use "machine learning stuff" to predict when this will happen and prevent it [Pat, CEO]. Yet, Chris suggested that further functionality would be useful for health researchers. This included individual-level metrics that "track people over time" and the ability explore relationships between engagement, acceptability and effectiveness:

> "I really need to tie it to one individual to see...was that usage related to what they thought of the website and how they changed their behaviour". [Chris, Health Researcher]

Overall, logging software was felt to help researchers collect user log data and to contain innovative features to assist with its interpretation. However, the interviews suggest that: setting up data collection requires input from multiple disciplines; the size and type of data it produces can be difficult for health researchers to analyse and interpret; and the logging software may lack some of the functionality required to fully meet health researchers' needs.

6.2.4 Assessing acceptability through device generated user logs

The use of logs to assess acceptability (i.e. participants' perceptions and experiences of the app) may increase efficiency through reducing the need to collect qualitative interview data. Yet the scoping review revealed that most studies still used qualitative methods to assess acceptability. Two themes

emerged from interviews in relation to this finding: Comparison of qualitative and logging methods in exploring acceptability, and the relationship between acceptability and engagement.

6.2.4.1 Comparison of qualitative and logging methods in exploring acceptability

Many participants highlighted the relative strengths and weaknesses of logs and qualitative interviews for understanding acceptability. Aaron [Industry Data Scientist] described how his company found interviews useful for exploring aspects of acceptability that are "difficult to quantify" such as "trust" and "transparency". Participants also felt that qualitative methods had the added benefit of helping researchers understand how to improve acceptability. Tom [Health Researcher] felt that using qualitative methods at the end of a study provided a "useful" opportunity to get "input" from users on how the device could more easily fit within their daily life and "workflow". Similarly, Steve [CEO] felt that logs were suitable if there was no desire to "change" the app (i.e. it was stable), however interviews were highly "valuable" in providing ideas from users that could inform or "guide" improvements to the app and additional app features:

> I think if I had a product that I didn't want to change, then I would be happy to just try to log data... I think we are trying to do a lot of things that aren't in the app, and that we need to try and understand before we build it, and like, kind of, as a guide to where to head next? So then I think that the interviews are very valuable. [Steve, CEO]

However, although most interviewees recognised the value of qualitative methods, the challenges of using them were also recognised. For example, Aaron [Industry Data Scientist] described how his company found it difficult to achieve the necessary face-to-face contact for interviews: "the scale that we're operating at, we don't have the ability to go out in close contact with everybody using the app". He noted that this might result in self-selection biases where only those motivated to use the app are interviewed:

> So, for us, for example, one of the problems is that anybody that we manage to convince to come in and talk to us.... there's going to be a certain amount of self-selection by us or maybe they'll be people who

> are ...nerds... or whatever you want to call them. So they may not be representative of the average user [Aaron, Industry Data Scientist].

Interviewing only "nerds" rather than "everybody using the app" may mean that overall results are less generalizable to other users. Aaron noted that log data, which does not require this face-to-face contact, enabled "looking across geographies" and to explore aspects of acceptability "more empirically". Overall, participants felt that qualitative face-to-face methods and logs had different strengths and weaknesses, and their suitability was dependent on the aim of the research, the stability of the product and the number of users.

6.2.4.2 The relationship between acceptability and engagement

Reflecting on the scoping review finding that few studies used logs to assess acceptability, some participants initially stated that logs showing engagement behaviours were "clearly a good proxy" [Liz, HCI Researcher] of acceptability, and that "acceptability could be interpreted as well, how frequently do they use it? Because if they don't use it at all then they are not going to be finding this intervention acceptable" [Maria, Health Researcher]. Nevertheless, participants went on to explain the data provided by logs should be "recognised" as only a proxy rather than direct measure of acceptability. For example, Liz [HCI Researcher] pointed out that "people may appear to be using a device for multiple reasons, not all of which are because they like it or it's acceptable to them". Tom, a health researcher, similarly noted, "there are certainly people who will use the device religiously yet not like various aspects of it" [Tom, Health Researcher]. Maria conversely provided examples for why users may accept a device but not frequently engage with it:

> But I don't think it is as straightforward a relationship as that… it might be that they find it acceptable, but they just don't see any need for them to use it that frequently, or they've used it once and they found out what it scores, they found out how they compared to other people and that was enough for them to want to change their behaviour and therefore they didn't need to log it in [Maria, Health Researcher]

The relationship between engagement and acceptability therefore appears to be complex, as acceptability is not the only determinant of engagement. Other interviewees described specific factors beyond acceptability that would help to

explain engagement patterns. These included: users' daily routines and "habits" such as "other apps that they use" [Aaron, Industry Data Scientist]; whether a user was only using an app frequently because they were aware they were in a study [Mark, Product Designer] and; how users behaved in their daily lives when the device "wasn't being used" [Tom, Health Researcher]. Importantly, these interviewees all felt that these factors could not be captured by logs.

In support of the use of logs to assess acceptability, Maria stated that there was likely to be "overlap" with the construct of engagement (i.e. they shared similar dimensions). However, Maria felt hesitant to rely on logs entirely, and feared omitting important data. Maria felt that further research was needed to fully understand and define the relationship between engagement and acceptability, and thus support the development of the standardised definitions of engagement and acceptability that are essential to facilitate comparison between studies:

> there isn't really that definition of acceptability, and it always seems to be slightly different defined by depending on who wants to know. ... And I think some kind of a shared terminology across, what we mean by usage or engagement or acceptability, so that when, you know, in the conclusions of any paper or the implications you say this was feasible, this was an acceptable app, everyone knows exactly what you mean by that, there's not room for misinterpretation. [Maria, Health Researcher].

Overall, both researchers and industry professionals widely agreed that although user logs can describe how a user behaves and interacts with a device, they cannot fully explain why participants behave in particular ways. Mark, an industry professional, stated that "the data tells you what, it typically doesn't tell you why" and Gillian, an HCI researcher similarly noted that logs could only tell "one part of the story". Participants felt that for a comprehensive understanding of user behaviour, qualitative or self-report data should be combined with, or "supplement", logs [Tom, Health Researcher] to "make a judgement on what is actually happening" [Mark, Product Designer].

6.3 Further research designs and methods

6.3.1 Pragmatic trials

The scoping review simply used the category of RCTs to describe all studies with a randomised design, however some interviewees (three health researchers and one HCI researcher) had specifically chosen to conduct *pragmatic* RCTs, as opposed to the traditional lab-based RCT. Joseph described the main differences between these types of trials:

> So, we did... a pragmatic trial. So, by pragmatic, we mean, yes, we do ask people to do certain things, but we can't enforce it, because they're not in a lab setting, they are in their day to day life setting. So we can ask them to do certain things... at the end of the trial, ask them about what they've done, and what are their views of the different things, and assess those. But we can't control them 24 hours a day, as if they were in a setting in a lab, or, which some of those clinical trials often do. [Joseph, Health Researcher].

Joseph described how pragmatic trials require participants to use the device "in the environment where they actually function". Some participants had used app stores to facilitate pragmatic trials, and described their advantages and disadvantages. Maria explained that this approach enabled researchers to demonstrate that an app would be found amongst other apps, and users would download it of their own accord, as opposed to being instructed to do so for study purposes:

> I think if I were to do it again I would do it the same way because I think the value of having it, that people find it naturally on the App Store does mean that what we know about the trial is that, like you can show that it's not that... it doesn't work in the wild, like people do actually choose to download this. [Maria, Health Researcher].

Thus, app store facilitated pragmatic trials were seen to facilitate studies conducted "in the wild", i.e. in real world settings, and be reflective of real engagement (choosing to download) and effectiveness (whether it "work[s]") (i.e. improve external validity). James [CEO] speculated that new software to facilitate such trials could save costs and allow researchers to acquire sensor data from very large numbers of participants ("population level data"):

> I mean, you know about Apple's… ResearchKit programs...if you want to get population level data and you don't have literally millions of dollars' worth of research funding or tens of millions then this is a pretty great way... and the sensors, I think they're on the Apple watch, they're really good these days. [James, CEO].

These cost savings might include for example, buying many accelerometers for participants or the costs of meeting with participants in person (Volkova et al. 2016). Nevertheless, in addition to these advantages of app store-facilitated trials, participants anticipated challenges. Daniel [Industry Data Scientist] explained that he had previously considered a pragmatic RCT approach, however he had been concerned about its ability to meet ethical requirements for patients to give consent "in front of a medical doctor that would explain the trial", and the consequent liability risks "supposing... we push them to exercise, they had a heart attack and died, they attribute it to us and we get sued". Daniel therefore felt that, although perhaps suitable for evaluating apps which target "healthy individuals" to promote "better fitness or something", app store approaches were less appropriate for apps targeting clinical populations.

Another perceived challenge of app store-facilitated pragmatic trials was ensuring participants completed necessary trial procedures without discontinuing use of the app. James, when discussing ResearchKit, felt that "the opt-in rate is quite low", i.e. that participants would not necessarily permit researchers or industry professionals to access their sensor data. Maria noted that, although app stores were a "feasible recruitment method", it was important to design the app to reduce the chance of drop out early in the trial:

> …basically it was how few questions can be asked of the users and still get as much information about their alcohol consumption and their drinking patterns as possible... a lot of sort of work went into...paring down the registration process... just because the huge dropout rate basically at the beginning [Maria, Health Researcher]

Maria felt that the drop out was because there "was too much burden on the user" in completing questionnaires. She recounted how, in collaborating with industry professionals to develop the app, she had been advised to ensure participants benefit from providing information:

> all of the advice in there was that this is the point in which you could lose people and that you need to give them something, you know, it needs to sort of almost be a reciprocal relationship... And if the user has to spend five minutes putting information into an app and doesn't get anything back they then are very quickly, it puts you off, and probably won't get to the point which they receive anything". [Maria, Health Researcher]

Aaron, an industry data scientist, also described how he had had to consider this "reciprocal relationship" (i.e. an exchange that is beneficial to both researchers and users) when conducting an app store pragmatic trial of a physical activity and emotional regulation app. He explained that researchers in a previous lab-based study of the app focussed on 'algorithms' (e.g. accuracy in detecting physical activity) and "didn't think" of reciprocity, as it was more of a problem associated with real-world settings:

> ...the sort of HCI considerations; how do you use that data to give some value to your end user? ... they didn't think of any of those issues, they just were assessing whether the algorithm could go in the right direction, but obviously if you want to make this happen in practice then, if someone's going to download an app, they need to see some value in it for themselves... how can we create a user experience that fulfils people via a mixture of passive data collection and user reports. [Aaron, Industry Data Scientist]

Therefore, both health researchers and industry data scientists felt that pragmatic app store approaches had key strengths and weaknesses, and that it was important to consider user experience issues to minimise dropout. This included ensuring that users felt it was valuable and worthwhile to provide requested data.

6.3.2 'Staggered release' designs: evaluating and revising prototypes

Participants currently working in industry, or who had worked in that sector previously, described using agile and rapid evaluation approaches that had not been captured by the scoping review. One characteristic of these approaches included testing early prototypes (i.e. not yet completed, finalised and fully developed) by making these available to users (either via app store release or in person), and using the feedback to then build something "more solid" [George, Product Designer]. Another characteristic was making this prototype available

only to a “small sample” of the user population initially, then adapting it based on results, and releasing the adapted version to a larger user group. This created a “staggered release” or “roll out” approach, which was described by Mark, an industry professional, who had used it to evaluate the effectiveness of apps:

> [our] preferred method is to build in rigorous evaluation phases. Whether those are part of our roll out, where we will do an RCT or an A/B type split tests, early, with a small sample, learn something, adapt it, put it out to a larger part of the client population... so, these kind of staggered releases where if you have a total population within a client, you can stagger those roll outs over time and test and adapt within each one of those phases". [Mark, Product Designer]

Aaron, a data scientist in industry, described a similar approach:

> It's like, oh, we have an idea, we think it could work. So people are like, well, ship it to users and ship it to one per cent of users, ship it to five per cent of users, and if it works then we’ll wrap it up, or we’ll scale it up. If it doesn’t work, we’ll look at the data and think again of what we need to do.... [Aaron, Industry Data Scientist]

The “A/B tests” described by Mark involve participants being “split” and assigned to one version of an app or the other (King 2017). All industry professionals described using this research design; however, most had assessed outcomes relating to “marketing” [James, CEO] or engagement [Aaron, industry data scientist], as opposed to effectiveness. One other interviewee had conducted what he termed “rapid RCTs” (Tom, a health researcher). Tom described using rapid RCTs in industry settings because “in the private sector we just did everything a little faster”. He explained that they were “a fairly rapid sort of randomised control trials or pilot trials depending on the circumstances” including “a massive recruitment... as quickly as we possibly could...”, online screening, and then a “a randomised trial pretty quickly of the prototype that we had developed or that was in the data phase, against some control”.

Industry data scientists and product designers described the advantages of these staggered approaches. Aaron [Industry Data Scientist] felt that an A/B approach, in particular, “seems much more in tune” and manageable in the face of the highly “dynamic” nature of apps, rather than simply going ahead to release a full version to a large number of people via an app store release, as he had previously done in an academic project:

> we had our big press release day, and then within two weeks we had a ton of downloads, but within that first month we had some mad panic all over the place, and especially because this was a public release. So there were phones that the app didn't work on properly, there were some bugs in the app, there were all these little issues. Our server wasn't ready to cope with that kind of load, and so we were basically putting out fires left and right and, if I compare it now to my experience here at Company7, that would never happen here because we would never release a feature to everybody. [Aaron, Industry Data Scientist]

Mark [Product Designer] described further advantages of staggered approaches: they enable an app to be available to everyone only if it is "beneficial", and allow effectiveness and acceptability to be "actively assessed" during the release, as opposed to before or after. Overall staggered approaches were considered to have several advantages; furthermore, no disadvantages to staggered approaches emerged during the interviews.

6.3.3 The remote assessment of acceptability

Beyond using face-to-face qualitative methods to assess acceptability (as described previously), some participants described methods that enabled them to evaluate acceptability remotely. Researchers in health and HCI, and industry data scientists, felt there were opportunities in Ecological Momentary Assessment or "EMA", also known as Event Based Monitoring or experience sampling. This involves sending short questionnaires frequently to a participant's mobile device, for them to answer questions at specific time points and in their own context (as opposed to asking them to recall relevant answers retrospectively in a single questionnaire at the end of the study, which can introduce recall bias) (Shiffman et al. 2008). Aaron (Industry Data Scientist) and Liz [HCI Researcher] had used this approach previously, while others mentioned it as something useful ("event based monitoring... you can achieve really nice results with that one" [Fred HCI Researcher]) that researchers use this more often ("we should probably be doing more event based monitoring" [Tom health researcher]).

While those mentioning EMA methods were mainly researchers, industry professionals described using other methods to assess acceptability, including contacting existing app users through the app itself. Pat [CEO] used this

approach to invite users to participate in a "personal interview", while Steve [CEO] explained that his company contacted users to ask if they would agree to having their own interaction with the app being visually recorded, using through screen-capture video software:

> Sometimes it might be easier to watch a video to understand what's wrong, and so you'll see the raw data of where they actually press, because you get, like a kind of a better sense of when they hesitate [Steve, CEO].

Others working in industry reported using app store ratings and online "user reviews" to understand acceptability [James, CEO], and user comments and feedback that people provide (e.g., through email), and how these are logged by customer support [Steve, CEO][Daniel, Industry Data Scientist].

6.4 Discussion

This chapter explored the perceptions and experiences of health and HCI researchers, data scientists, and industry professionals in relation to key scoping review findings. The majority of studies evaluating physical activity apps and wearables identified in the scoping review used RCTs. The researchers and industry data scientists interviewed clearly perceived RCTs to be the 'gold standard' research design, and felt the expectations of academic journals and requirements of organisations ultimately adopting behaviour change apps and wearables were contributing to the continued use of RCTs. Interviewees across disciplines and sectors also acknowledged the accuracy of external research-grade sensors (as often used in RCTs), but felt that important factors that should inform decisions about using external or in-device sensors, such as reliability, user burden, engagement and data availability.

Most studies identified in the scoping review employed user logs to assess engagement. Interviewees valued the objectivity of log data and the availability of logging software, however health researchers experienced challenges using and interpreting these and felt more guidance was needed. Relatively few studies employed user logs to assess acceptability. Interviewees felt that qualitative methods offered additional advantages in informing app design improvement, but that user logs might be appropriate to assess acceptability if a product is stable. However, interviewees also felt a better understanding of the

complex relationship between engagement and acceptability was needed to inform the design and deployment of efficient data collection methods to assess acceptability. Interviewees also highlighted important disadvantages of looking at log data alone, such as their limited capacity to understand "why" users behaved the way they do. As discussed in the thesis literature review (Chapter 2), this has been recognised in the wider research community (Dumais et al., 2014, El-Nasr et al., 2013, Kwasnicka et al. 2015), and underscores the need, when possible, to use other data collection methods in combination with log data, as opposed to log data alone.

Interviewees used research designs and methods beyond those captured by the scoping review. Researchers distinguished between traditional and pragmatic RCTs, and both researchers and industry professionals described experiences and perceptions surrounding the use of app stores to support pragmatic trials. App stores were considered to save research costs and improve external validity, while reported challenges included acquiring consent from clinical patients and difficulties with user drop out. To reduce drop out, participants felt it was important to offer some form of value to users in participating in a trial. Industry professionals described other designs not included in the scoping review, including A/B testing and rapid RCTs, staggered release (i.e. agile) approaches, and innovative methods for assessing acceptability remotely.

Some participants identified with more than one discipline, which is perhaps not surprising given the interdisciplinary nature of mobile health research. HCI research has itself been identified as a highly interdisciplinary field (Blackwell 2015), and defining "data science" as a discipline or specialist area also be challenging (Provost and Fawcett 2013). Describing individual disciplines and their fundamental principles can increase knowledge surrounding what experts have "to offer" (Provost and Fawcett 2013, p.51) and improve interdisciplinary collaborations (Pagliari 2007). However, those who have experienced working across disciplines and sectors (and their ability to wear different "hats") can be valuable informants. For example, reflecting on their experiences in industry, interviewees described how things were done "faster" or advantages of using staggered approaches over the full release approaches often used in academia. Thus focusing on these individuals' experiences and perspectives could not only

advance understandings of disciplines' strengths and weaknesses in evaluating mHealth (as recently reviewed by (Blandford et al. 2018), but guide efforts towards increasing research efficiency and rigour.

The interviews did not suggest there were stark contrasts and opposing perspectives between disciplines. However, health researchers in particular described challenges in using logging software to assess engagement, including: needing assistance with setting it up; challenges interpreting an overwhelming amount of data collected; and felt more guidance was needed on what engagement and acceptability measures to use. The fact that health researchers required other colleagues to assist them highlights the need for interdisciplinarity within a project (Nilsen et al. 2012). Importantly, interviewees described using freely available tools such as Google and Apple analytics, however these are not necessarily tailored to health research. New health-centered analytic tools are becoming available (University of Southampton 2016, Morrison and Doherty 2014), but more research is needed to review and/or (further) develop free and comprehensive software for health behaviour change researchers. This should take into account some of the needs identified in this chapter, such as longitudinal individual-level metrics and the ability to explore relationships between engagement, acceptability and effectiveness.

Industry professionals used evaluation approaches characterised by "staggered release" of a prototype to only a small portion of users, with continuous testing and adaptation, and repeating this process while continuously increasing the number of users to which the app is "rolled out". This approach is essentially "agile". Hekler et al (2016) describe an agile process for developing and evaluating health behaviour change technologies that may increase efficiency, but does not incorporate a staggered release approach for assessing publicly available apps. Interviewees described further advantages beyond efficiency, such as ensuring a product is effective for a small representative sample before being made available to large numbers of users; however, more research is needed to understand the strengths and weaknesses of this approach.

It will be important to explore how staggered A/B tests, in particular, differ to RCTs in their experimental set up, efficiency, and contribution to knowledge. Both research designs involve randomly assigning individuals to different groups. A/B designs are typically used for design-related outcomes (such as comparing user response rates between two groups receiving webpages with different layouts, King et al., 2017), and as such can produce rapid results. RCTs, on the other hand, can require assessing individuals at two or more time points, and measuring long-term outcomes. While more time-consuming, these RCT features promote rigour and produce useful results (such as understanding effectiveness over time). More research is needed to explore how and when A/B tests can be most usefully applied within app development and evaluation cycles.

Other themes revealed intricacies and differences in perspectives within disciplines; HCI researchers differed in whether they perceived RCTs to be suitable and useful for HCI research into behaviour change apps and wearables. This finding reflects a wider on-going debate in the HCI community on the importance and suitability of RCTs for HCI research (Klasnja et al. 2011, Stawarz and Cox 2015). Experiences within disciplines are also likely to differ: some HCI researchers had participated in conducting RCTs and others had not, which may have influenced their perspectives. The interview findings could be followed up using large-scale surveys with appropriate representative sampling reflecting particular disciplines and sectors. Results could determine whether it is worth investing in tailored solutions and support for particular disciplines in evaluating mHealth (such as health researchers in using logging software). However, the interview findings suggest such questionnaire studies should be designed carefully. “Hybrid” researchers may not identify with a single discipline, and there are risks of oversimplification in characterising perspectives according to disciplines and sectors that are instead more dependent on individual experiences (such as conducting RCTs).

A complimentary and fruitful approach to reviewing and developing logging software for health researchers would be to develop guidelines for the evaluation of behaviour change (including physical activity) apps and wearables. Guidelines could increase efficiency by reducing the overall time and effort researchers spend trying to interpret an overwhelming amount of data, requiring

fewer people (i.e. from multidisciplinary teams), and creating standardised measures that can enable results to be compared across studies and thus rapidly advance overall knowledge. Guidelines could further address which sensors (external or in-device) to use to assess physical activity outcomes. The interviews suggest that although external research-grade sensors are gold standard, practical and logistical factors, such as user burden, should also be considered. These may be particularly important for pragmatic trials.

The interview findings also have implications for scoping review methodology. Future reviews in this area should categorise studies according to whether they used “pragmatic” RCTs or standard lab-based RCTs (i.e. whether they took place in research or real world settings). This will also give an indication of how many researchers still use “lab-based” non-pragmatic designs relative to pragmatic designs, to evaluate physical activity apps. This may be an important distinction as different types of RCT may differ in their suitability for evaluating mHealth technologies.

may differ in their suitability for evaluating mHealth technologies. One reason for consulting with experts on scoping review findings is to identify any important issues or items that had been omitted (Arksey and O'Malley 2005, Levac et al. 2010). Methods such as app store and agile approaches that were not captured by the review may be important means of increasing research efficiency. To learn more about these and their appropriateness for assessing the impact of physical activity apps and wearables, scoping reviews should include any research conducted or reported by industry professionals within their search strategy.

6.5 Conclusion

Interviewees’ experiences and perspectives provide further interpretations and possible explanations for the scoping review findings. However, the results reported in this chapter often reflect the researchers’ and industry data scientists’ experiences, perceptions and practices, rather than those of industry professionals’. This meant little was revealed about industry professionals’ perceptions and experiences relating to evaluation in general, which is further explored in chapter 7. The current chapter also discovered possible factors that

may result in researchers continuing to use RCTs. The next chapter, 7, will more closely examine factors associated with the limited use of rapid research designs, specifically.

Finally, the current chapter reported interviewees' experiences and perspectives on using app store approaches for assessing effectiveness. One advantage reported is being able to understand whether the app is effective in real world settings, and (as discussed in the Literature Review) this approach may also improve research efficiency. The next chapter reports how automation, specifically, can increase efficiency, and automated app store approaches for physical activity apps are explored in chapters 8 and 9.

Chapter 7 Encouraging the evaluation of effectiveness and the use of rapid research designs

7.1 Introduction

The previous chapter described the current practices of researchers, industry data scientists and other industry professionals in relation to the scoping review findings. Two issues suggested further exploration of the interviews was needed. Firstly, little was revealed about industry professionals' perceptions and experiences relating to evaluation, as it was mostly researchers and industry data scientists who described relevant experiences, perceptions and practices relating to many of the key scoping review findings. Secondly, chapter 6 reported some perceptions and experiences that may help to explain why RCTs, specifically, continue to be used by researchers involved in mHealth projects; these may only partially help to explain why rapid research designs are not adopted.

This chapter reports further analysis of the interviews. The COM-B model was used to identify barriers and facilitating factors which influence: firstly, the evaluation of apps and wearables targeting physical activity and other health behaviours, and second, the use of rapid research designs to do so across both academia and industry

In brief, COM-B model proposes Capability, Opportunity and Motivational factors that may promote or prevent engagement in Behaviour, and these model components each have two associated subcomponents (for a full description of the COM-B model, see Chapter 4, Methods, 4.5.9).

7.2 Motivation for evaluation

7.2.1 Reflective motivation

7.2.1.1 Interests beyond effectiveness

A key emergent finding was industry professional's lack of interest and motivation to assess effectiveness. Industry professionals instead appeared to be primarily driven by financial gain:

> ...businesses evaluate how much money they're making, you know? I used to be in academia and the people that are evaluating, I don't know, effectiveness and things, but at the end of the day... for most companies it comes down to money... it is important for academics and researchers to know whether these things work, but that's not necessarily actually of interest to the people making the apps. *[James, CEO]*

> I do think it actually comes down to the difference between organisations which are founded on a principle that they're trying to achieve change, and other organisations which are founded on the principle of trying to make money... Financial organisations are there, primarily, to make money. [George, Product Designer]

Thus, industry professionals appeared not to associate effectiveness evaluations with financial gain. On the other hand, they strongly felt that user engagement and "active users" were more indicative of financial gain than effectiveness, and so more of a priority to assess:

> money... a proxy for that is engagement and it's also active users.... As long as people are buying it and they're running and they're opening the app and they're telling their friends it's good and they're rating it on the app store ... the experiments that you will run if you do run any experiments will be how do we monetise people better, and how do we improve engagement [James, CEO]

James felt that although his company wanted users to "tell their friends and keep on playing", this was not "the same, unfortunately, as... effectiveness... it's a different sort of thing". Another industry professional, George, emphasised that industry professionals tend to be interested in "desirability" (whether people "want" the app), feasibility (i.e. technologically and "functionally, it works") and viability (i.e. economically; "can we make

money"). He described these as fundamental and "baseline" concerns for designers, but that effectiveness was something "additional".

There were mixed views amongst HCI researchers on the importance of effectiveness. Some were interested in using technology to support people in the stages before behaviour change (e.g. in raising awareness) or afterwards in maintaining a behaviour:

> It's not about behaviour change, usually not.... most people are in phases outside of behaviour change and I believe that it's much more important to support these phases of relative stability… raising awareness, very, very small steps until somebody ultimately really undergoes actual behaviour change. [Fred, HCI Researcher].

> It's not always about changing behaviours, it's sometimes just about doing what, continuing to do what you've been doing for some time. Or adapting behaviour as circumstances and needs, situations change. [Liz, HCI Researcher]

Others, however, felt there was a growing expectation from funders for HCI researchers to evaluate effectiveness (i.e. "impact"):

> I think it is still new for the HCI community... but it is inevitable because... research-funding bodies expect researchers to talk about impact. I think it's inevitable to kind of step alongside a lab study and do larger scale user studies including RCTs for an HCI researcher [Gillian, HCI]

Interestingly, researchers across disciplines appeared to share an interest beyond effectiveness that centred on implementation and evaluation in real world settings. HCI researchers believed the HCI community had shifted their attention to, in particular, the "broader context" of device use and engagement [Liz, HCI Researcher]. This involved exploring "what people do outside the lab" [Gillian, HCI Researcher] and how technologies "fit" into people's lives:

> It was more the broad, how the people fit these technologies into their lives, what are the challenges, what are the good features of some technologies that other ones don't have... I've seen it in a lot of researchers... it's not always just the details of the interaction design that matter, it's also how stuff fits within the broader context of use.... And what makes technology acceptable, and fit for purpose, in that sense. [Liz, HCI Researcher]

Michael, another HCI researcher noted how this focus on users' contexts could make an app more effective or supportive in changing behaviour:

> you really need to dive into the context to see what empowers people to keep on using the tool and keep on engaging with the tool and then to, yes, then to make sure you support them as much as possible to actually change their behaviour [Michael, HCI Researcher]

Health researchers all conducted effectiveness evaluations and assessed impact but were also interested in reach, dissemination and implementation. Tom felt that as well as, or instead of, assessing impact studies should examine reach and scale:

> I see way too many evaluations of mobile or digital interventions. People didn't build the digital intervention for it to be necessarily better than the in-person sort of equivalent. They built it to just have greater reach. They didn't even really care if it was a little less effective than the in person but they wanted to have the reach and scalability yet what they ended up doing in the study was evaluating its efficacy [Tom, Health Researcher]

Another health researcher, Chris, felt strongly that "we do randomised control trials all the time" and too few studies consider the "next step" towards implementation and dissemination:

> you see almost no research taking the next step doing a dissemination study, what are the methods that we need to use to actually expose this to a lot of people so that a lot of people actually start using it... let's spend a fair bit of money, just on figuring out how do we actually get people's attention... more implementation science really [Chris, Health Researcher]

Chris therefore felt that more research, and more funding, should be allocated towards assessing implementation and understanding how to increase intervention reach. He further explained that getting "people's attention" was especially important for behaviour change technologies because the Internet and app stores are "crowded" with other websites and apps. Maria, another mHealth researcher, felt that researchers' next steps beyond effectiveness included considering further intervention development, optimisation, implementation and that "more evaluations that lead onto further development" were needed:

> We... get to a product and then evaluate it and then it's like okay, job done. And I think there's a huge amount of work to be done with this for academics that sort of designing an app or a digital product or whatever it is with a long-term plan in mind. It's whether you want to implement that in the NHS or they sort of have to how they impact or probably, you know, sort of big picture things, you need to know at the beginning what those stakeholders would want for it. [Maria, Health Researcher].

In addition to HCI and health researchers, an industry data scientist reported focussing on real world settings within his work when in academia, as opposed to those in "very controlled environments". He felt that the data becomes "interesting to work with for a variety of reasons" in real world settings. He recounted that "the data was much sparser than we expected it to be" (e.g. missing data issues) and that "there were some impossibly high values in the data, as if they were on a rocket going to Mars or something, so we clearly had to clean that" [Aaron, Industry Data Scientist].

Overall, industry professionals' ultimate motive to make a profit and to improve and maintain users' engagement with their product competed with, and acted as a barrier towards, assessing effectiveness. Health, HCI and data scientists also felt constructs beyond effectiveness were important, and shared similar interests in evaluations taking place in in real world contexts.

7.2.1.2 Company reputation; the potential for "good press"

Industry professionals felt there were possible marketing opportunities and the opportunity to receive favorable media coverage if they assessed effectiveness. James noted it could be a "a great marketing thing... if we're able to prove, I suppose, that it worked then that would be a really good press release (James, CEO)". However, the data that could be used in effectiveness studies was not seen as a "gold mine" for the company [James, CEO], and "good press" or a favourable press release (i.e. "PR") alone were not considered likely to make much of a difference to the company:

> We have theoretically a huge amount of data...it's a bit of a goldmine, and we just haven't really done anything with it…well, it's not a goldmine for us, it's probably a goldmine for someone else, I

> mean... we haven't really looked because in some ways, what difference does it make? (James, CEO)

As further indication that media coverage and marketing opportunities alone may not be powerful enough motivators to assess effectiveness, another industry professional recounted a time when a company he had been working with had been interested in marketing opportunities, yet the effectiveness evaluation was not completed. George felt this was because ultimately they were not focused on behaviour change:

> "in that particular incident, the client was more interested in an app as a marketing thing, than a valuable change tool.... So, some organisations are evidence based through and through... Others are less interested in it, it's more a marketing construct. Although they might pay lip service to it, they're not particularly focused on it." [George, Product Designer]

Overall, although the opportunity to improve company reputation was a potential facilitator for motivating industry professionals to evaluate effectiveness, they did not see a clear link between this and making profit.

7.2.1.3 Perceived risks

Despite the opportunity for a company to improve their reputation by assessing effectiveness, interviewees highlighted the associated risk of a company's app being found to be ineffective:

> on the positive side you can really make name as a developer of stuff that actually really works, as proven by science. But, on the other hand, you can also find out that maybe it doesn't really do what you expected it to and then, yes.... there's the risk of being disproven [Michael, HCI Researcher]

An industry professional explained how he was unsure whether and how negative effectiveness results would affect a company, suggesting that on the one hand there "may be a deadening of trust" and that "from a commercial perspective, I suppose they just stop it, stop making money from that thing", but on the other hand other industries such as diet programs had not suffered that fate:

> Think of the diet industry, all the kind of like endless fatty, sugary, diet replacement meal drink things, that have been proven time and

> time again, not to actually help people lose weight, and they're still going strong [George, Product Designer]

People from industry associated effectiveness evaluations of health related products with greater risk than engagement assessments for other technologies. They felt health was a was more challenging and intimidating area or domain than, for example, evaluating social networking sites:

> I guess in the tech industry, or at least in some segments of it, the biggest difference is that there's maybe slightly less risk than in a very strict health area... you know, the main risk for us... making a significant product change, that this kills the user experience for people [Aaron, industry data scientist]

George highlighted other risks associated with liability and the importance of acquiring the correct "permissions":

> "if you're recruiting people after they've downloaded something, you need to be really careful to make sure that you've got proper permissions.... Because, actually, there's an awful lot of very sensitive data that's gonna be handed over." [George, Product Designer].

Overall, those who worked in industry felt their company faced risks if they developed and tested products that were found to be ineffective, and that these risks were higher in health related domains.

7.2.1.4 Morals and social good

Another factor that could facilitate industry professionals' motivation to evaluate behaviour change was the concept of innate 'morals'; some described being motivated to do "good". However, in the same way that industry professionals weighed up whether "good press" would be financially worthwhile, they felt companies ultimately consider financial sustainability over morals:

> I'm interested in technology for good... There seems quite a big moral dimension to this... when I'm working in commercial organisations I've not got inner thoughts, oh my god, they're all immoral - immoral bastards, bleeding people out of every penny that they possibly can. But there is something subtle that kind of happens in the culture where you are more focused around how much income is coming in every week, rather than, are you achieving your goal of trying to make change in the world. [George, Product Designer]

One industry professional felt that health apps that modify behaviours could possibly be harmful or "dangerous" to users and as such, they should be evaluated for effectiveness. George suggested studies should be conducted to provide transparency to users (e.g. an understanding of the true likelihood of successfully changing behaviour):

> if you're starting something to help people change their behaviour and it doesn't, that's quite dangerous, all of a sudden... or they don't know whether their app has worked or not, at all. That seems like a dangerous position to be in.... I can't think of a specific example where it might put people in danger, but I mean, I presume that could happen, if you were kind of like tinkering about with some behaviours... there should be a kind of transparency.... for users, if you're saying this is going to do something - how does it do it, has it done it in the past, what's the likelihood of success? [George, Product Designer]

Overall, motivation (or lack thereof) for evaluating effectiveness was a key theme for industry professionals in particular. Mark reflected on his experiences in designing and evaluating behaviour change apps for other industry professionals (i.e. "clients"):

> It really comes down to... does our client partner see the value in doing it [evaluating effectiveness] at that time? Their motives may be, we need to release something, or they are still more interested in, well, do we get press from this, is it about downloads, is it about use, is it about, you know, the story that we can tell that we have designed something, right, or put something out in the world? [Mark, Product Designer]

7.2.2 Automatic motivation

No factors were found that related to automatic motivation for evaluation.

7.3 Opportunity for evaluation

While motivational factors were described for different disciplines in evaluating the effectiveness of health apps and wearables, the remainder of the analysis focuses on Opportunity and (subsequently) Capability factors in evaluating effectiveness with a specific focus on industry. The COM-B model proposes that

an individual's behaviour can be helped or hindered by their environment via both physical and social opportunities.

7.3.1 Physical opportunity

7.3.1.1 Time-related factors

Interviewees perceived that effectiveness evaluations to take a long time to conduct. All interviewees described how the overall evaluation process and "science" in general can develop at a "glacial pace" [Michael, HCI Researcher]. This included conducting reviews, designing and conducting pilot and full experiments, long data collection and "testing periods" [George, Product Designer], and publishing:

> Yeah, industry time... they want solutions and they want them quick. You know how researchers do, they, they start a review, then they start the design, then they do a pre-test, then they do another test, and it all needs to be tightly controlled and measured and all the controls right. By the time they get it going, they've [industry professionals] moved on to something else altogether...They don't have that kind of time, they want a solution in six months or something.... Those timelines won't work for researchers. [Chris, Health research]

The duration of evaluations was considered to be incompatible with industry professionals' time scales, and thus a barrier to them evaluating effectiveness. Specifically, the rapid "prototyping" aspect of industry professionals' practices in designing apps was felt to contrast with effectiveness evaluation timings [George, Product Designer]. Michael, an HCI professional who frequently worked with industry professionals, highlighted that rapid development procedures are also driven by the need for rapid investment from funders (i.e. as further described in "social opportunity", below).

Time-based barriers were highly linked with financial barriers and other factors within the COM-B framework. Michael noted that to pause or "freeze" prototyping and development work to facilitate lengthy evaluations was not financially viable for industry professionals:

> ... if you have to freeze your products for three years waiting for the scientific results to come in, then it's going to be technically obsolete too... by the time results are published, they have already done three iterations to their products and the thing that is actually on the market is not exactly the thing that has been tested... they can't afford to freeze it. They just don't. [Michael, HCI Researcher]

The statement that "they just don't." emphasises time as a barrier to evaluating effectiveness for industry professionals. There were strong associations between the perceived time it took to conduct effectiveness evaluations and industry professionals' motivation to evaluate their products. The usefulness of the research to them was hampered by the large amount of time evaluations required:

> whether health and fitness apps work or not.... honestly part of the issue is just fundamental to academia…because people say, oh, is this research helpful to developers? And it just can't be because the research takes years to come out, and the app development world moves so quickly and the technology moves so quickly that… I saw there was a study that came out about Project3 and this was about Project3 from three years ago and it's not that long ago, really, but…the app's changed a lot and it's improved a lot, and so I think that's a real challenge, and I don't know how to solve that. [James industry professional].

Some participants highlighted the "automation" of trials as a potential facilitator for industry professionals' effectiveness evaluations, by reducing their overall duration. One product designer, George, reflecting on his involvement in a project developing a text-message based behaviour change intervention, felt automation would have been useful to assess long-term outcomes in particular:

> we kind of knew that it would be a limited prototype... that we could probably get people to take three days off a week, for three out of four weeks in a month. But that doesn't really tell us whether that achieves any type of longer behavioural change, over time. So we needed, still, something that could automate that, and look at the data over a patch of three months, six months, a year, and kind of like, just do some observational to see if it's working longer term [George, Product Designer].

George's use of the term "observational" suggests this automation process would require minimal input from those investigating effectiveness, and simply collect data over time. Tom, an health researcher, went further to suggest

experimental evaluations could be "embed[ded]" into the development and early releases of the app:

> I just think anything that we put in there that will slow down the process for them potentially will be something they won't want to do. If you can embed evaluation into the development and launch and commercialisation process for them I think that makes a lot more sense [Tom, Health Researcher]

Tom's quote suggests that because slow evaluations may demotivate for industry professionals, speeding these up via embedded evaluation may be an appropriate approach. Tom went on to describe the characteristics of automated trials, including 'real time' analysis and evaluation during dissemination:

> I mean, I keep hoping we'll get to this point.... say you developed a smartphone app or whatever and instead of doing... a formal evaluation of it, you just, you know, you commercialise it. You put it out in the field.... and you put it out in two versions.... that's the two arm RCT version of it. You could do other things. The outcome data is collected by the device itself. Then that data is coming back to you in real time... we can evaluate them as they're being disseminated. That's I think where the real promise is in doing this and cuts through having to spend five years doing a formal two arm RCT [Tom, Health Researcher]

Thus Tom felt automation was key in reducing the duration of evaluations such as RCTs. James also felt that industry professionals would be more motivated to participate in evaluations of their products if they did not take up much of their time. He described an automation approach whereby evaluators (e.g. researchers) would provide industry professionals with a template to embed in their product that automatically collects certain types of data. However, he explained that researchers would need to ensure it was integrated into development and in ways that would not negatively impact the product or company:

> I think they want to help but they just don't want to spend a lot of time doing it, and so if there was somebody who said, well...here's an XML schema [automatic data collection tool] or whatever, then I think people would be interested in that, but yes, I think that there would have to be a lot of hard work done to make sure it was smooth and kind of bullet proof as possible. [James, CEO]

Overall, the length of time evaluations take was a barrier to industry professional in evaluating effectiveness, which was linked to other opportunity factors and appeared to have a negative influence on motivation. Automation of evaluations was perceived as an opportunity to speed up the evaluation process (and thus a facilitating factor).

7.3.1.2 Monetary and staff resources

Industry professionals perceived effectiveness evaluations to be costly. They felt that they did not have sufficient funds for evaluating effectiveness as "budgets are really tight... that's kind of like the commercial reality" [George, Product Designer]. Another industry professional (Mark, who provided development and evaluation services to other companies) felt that industry professionals were likely to prioritise and allocate resources to developing an app, improving it (in terms of user experience), and releasing it, rather than evaluating effectiveness:

> an evaluation costs money, you know, and time, but especially money. If they have a budget and they do, where do they want to put most of their resources and finances? It typically goes to building the best product that they can and getting it out [Mark, Product Designer]

Interestingly, it was especially "rigorous" experimental methods [James, CEO] that interviewees felt they could not afford as this required more than simply comparing "user reviews" [George, Product Designer]. Industry professionals perceived bigger, more "mature" [George, Product Designer] tech companies (in non-health domains, such as social media, music and gaming) to have greater resources than themselves. They felt this allowed them to conduct rigorous experiments, and specifically, to employ data scientists and academics to do so:

> Facebook famously does all sorts of experiments on its users, and they have very smart and experienced data scientists and actual academics who design their experiments... and of course if you're Google or you're Facebook or whatever, then you can have a user base in the billions and you can set up 1000 experiments, changing the variables and so on [James, CEO]

> the [experiments] that you describe I would assume is more like... the music, like Spotify, and the big gaming companies, because you need quite a big data team... at our company there's five people in the Data Team....We would need to regrow the team quite a bit.... so I

> would say that we are at least months or years away from moving into that. [Steve, CEO]

Steve went on to explain that the few data scientists currently employed in his company were busy maintaining large data sets and "infrastructure" relating to marketing and developing the product, and so more data scientists would be needed to evaluate effectiveness. Interestingly, an industry data scientist explained that "multiple disciplines" (i.e. such as "designers" and "engineers") were needed for the "pieces of puzzle to come together", to facilitate a trial. Overall, there was perceived to be a need for multiple staff to conduct effectiveness evaluations, including, but not limited to, data scientists. Interviewees felt they did not have the resources to employ these staff.

7.3.2 Social opportunity for evaluation

7.3.2.1 Funders and financial decision-makers

While industry professionals felt that they did not have the finances to conduct experiments, a related major barrier to evaluating effectiveness appeared to be the motives of external organisations that provided funding (such as commissioners and investors), and internal staff who made funding-related decisions (such as higher managerial staff).

Mark, an industry professional, noted that app evaluations within industry are typically not government funded or paid for "by the Chief Health Officer or things like that". Michael, an HCI researcher, described effectiveness evaluations as "a lot of work that nobody pays you for" in industry and, as such, is work that companies would have to "do in their own time". He felt that commissioners are not always interested in effectiveness, and unlikely to fund effectiveness (or "validation") studies. Michael also highlighted how commercial "investment agencies", who provide financial backing to digital start-up companies, want a quick "return on investment" in terms of a product being developed quickly and earning money (e.g. via downloads):

> the speed in which they need a return on investment, that is absolutely a challenge... what usually happens with start-ups is that they need funding and for that they need rapid growth...and for rapid

> growth you need really fast iteration cycles, then it becomes impossible to rigorously validate [Michael, HCI Researcher]

Mark, an industry professional, similarly reported that when projects are funded "they are really only interested in downloads and things like that". Michael described how this external pressure to work quickly and focus on app development and engagement becomes "a bit of a culture" where developers "could be going slower, but they don't dare to". A participant with experience of working in different sectors spoke about how this internal "culture" focusing on finance as opposed to effectiveness is driven by managerial staff and "cascades down" through the company:

> you see it really, really clearly, I think, when you sit in board meetings... when I used to work in a big commercial agency... the first agenda point, nearly always, is finance... where are we against our projections, are we doing well, are we in trouble, are we kind of like in the sunshine, are we making hay... in charitable organisations, finance is often the last question. And the first question will be, how are we doing, how are our programmes working. And I think it just cascades down from the top... it's such an engrained cultural thing [George, Product Designer].

The companies where data scientists worked, on the other hand, were more supportive of experimentation and evaluation. Daniel, who worked in a research company and evaluated app effectiveness while employed there, described how he was told by managers to study "whatever is scientifically interesting", and Aaron, another industry data scientist, noted how staff were encouraged to use experimental methods to assess app engagement. He described how a "mantra" -"design like you're right but test it like you're wrong" operated in his company, through the CEO having a vision to support rigorous experimentation to assess engagement, which was then "translated" by staff into empirical measures as "you move down the ranks":

> I guess, in practice, from a very high level, the CEO level, they're there to provide us with a vision and then that vision gets translated as you move down the ranks and then it gets translated into measurable outcomes... we have some internal tooling that allows us to set up these A/B tests, and those same tools then report these metrics and the guy who's our head of experimentation has done a fantastic job on this because it even does the significance testing and stuff like that.... [Aaron, industry data scientist].

Hence, although in the context of engagement, Aaron's company socially supported experimentation to the extent that managers were hired specifically for experimentation, and software was created to support data collection and analysis. Overall, some interviewees had experienced supportive workplaces in relation to evaluation, while others had not.

7.3.2.2 Competitors

In addition to managers influencing whether effectiveness was evaluated, there were also wider social influences in the form of industry professionals' competitors. Mark, an industry product designer, who provided app development and effectiveness evaluation services to other companies, described his experiences in motivating clients to conduct evaluations. He noted that while his own company had "strong motives" to evaluate effectiveness it sometimes involved "selling a little bit on the client end" (i.e. persuading other companies that they should evaluate effectiveness). He reported that momentum builds as competitors learn that others are evaluating apps:

> it took a while to build up the arguments and the incentives to get clients to do that and then as you start to get one, then it gets a little bit easier to do the second one and the third one... like, oh, you're doing this for like five other people and some of those are competitors. Yes, we want to do that too. (Mark, Product Designer).

As evaluations in industry are still rare, Mark described how he attempted to motivate clients by informing them that evaluations would be "a huge differentiator" and set them apart from "a hundred diabetes apps", of which few will "work".

Relatedly, industry professionals anticipated that if other companies did conduct evaluations they may bias their results to improve their reputation ("if you're marking your own homework, the temptation is to look for all the things that show that it's working") [George, Product Designer]. Another industry professional described such an approach as part of the "the sad, sad, nature of the business". Therefore there was a view that effectiveness evaluations should be conducted by impartial, knowledgeable evaluators with "expertise" externally "brought in":

> … industry, it's very interested in looking good and telling their story about this thing works... someone with very crude, knowledge base and, you know, statistical skills or understanding on studies would look at that you would say, like this is junk, right, there's no foundation here... so the importance of having maybe an outside evaluator or, you know, an academic partner or an impartial reviewer, whatever that is, especially with commercial services I think, you know, I think it helps you move that bias and that's important [Mark, Product Designer]

Overall, although industry professionals were not themselves entirely motived to evaluate their app's effectiveness despite the potential for "good press" (see section 7.2.1.2), they recognised its importance for ensuring competitors were truthful and not at an unfair advantage.

7.4 Capability in relation to evaluating effectiveness

The COM-B model includes two subcomponents of capability: psychological capability (the individual's ability to engage in necessary thought processes) and physical capability.

7.4.1 Psychological capability

7.4.1.1 Experimentation

Industry professionals perceived the procedures involved in conducting experiments to be challenging. James expressed "it's quite difficult to do in a scientific way". Evaluating multiple app features was seen as being particularly difficult. George [industry professional] described how evaluation became "exponentially more complicated" as the number of features grew. Another industry professional, Steve, similarly noted that a "crowded" app (i.e. with many app features) created evaluation difficulties. Interestingly, Steve indicated that because of this complexity he was less motivated to evaluate effectiveness and "can't be bothered":

> maybe it's a great feature if... we would have taken away three other ones, but when you add it on top of three that already exist, maybe it's just a pain in the ass, and you can't be bothered. So those are the things that are quite difficult to see how to test. [Steve, CEO]

Industry professionals also perceived there to be challenges in evaluating effectiveness due to real-world confounding factors such as seasons, the weather, and app users' motivation levels:

> I think obviously there are things that are tricky to measure... There's like a kind of huge seasonality as I said, New Year's resolutions, getting ready for the beach... after the holidays... Thanksgiving. It's like all these things that happen.... there's a huge difference, depend[ing] on if you start a diet on a Monday or a Saturday. So, how much does this prelude the data when you release a feature? So for example, if you want to have people exercise, if it's great weather people are more likely to go out and run. So what happens if there is shitty weather the same day as you release the new running feature? [Steve, CEO]

Rhetorically questioning "what happens" when there is a change in weather suggests that Steve felt uncertain as to how such confounds might be overcome or experimentally controlled. The potential for confounds appeared to negatively influence another industry professional's motivation to conduct experiments:

> I've done experiments and I've done studies, and I know that the data you get is just so noisy, and it's really hard to draw conclusions from that... Like, you know, I can look at, I don't know, how much do people run when they join Project3 over time. Maybe it goes up and maybe it goes down, but that doesn't necessarily tell me that Project3 is the cause of that. Maybe they were just interested in fitness. [James, CEO]

In addition to experimental procedures, James perceived challenges in framing experimental research questions ("you have to ask the right questions as well") and also in ensuring that results provide clear conclusions as to how to improve the app. He perceived that while ideally "I would hope" test results would suggest that "clearly you should do X, Y and Z [but] I know it's usually not that straightforward".

Overall, industry professionals felt uncertain about experimental procedures they felt could be useful or necessary (i.e. testing multiple app features, controlling for confounds). Along with recognising that experiments do not necessarily provide certainty and actionable outputs, these appeared to actively dampen any motivation to conduct effectiveness experiments.

7.4.1.2 Statistical analysis

While industry professionals themselves did not express difficulties in performing statistical analyses, other interviewees believed industry professionals had limited statistical skills. Aaron, an industry data scientist, felt that those developing apps [“product people”] often did not understand the concept of statistical significance:

> ...your typical product person won't necessarily have a strong statistical background... One of the mistakes that used to happen in different products was that people would be running an online A/B test and they see their key value go from .1 to .09 to .085, and so they’d say, look, this is going towards significance, let me just turn it off and say that it's successful. (Aaron data science).

Aaron’s account illustrates a mistaken conceptualisation in industry that things “trend towards significance”. Similarly, an HCI researcher told of her surprise in finding app developers who are highly experienced in collecting log data, often lack any analysis skills:

> And so they're collecting lots of stats... the stats from the student's study would be stored separately, and she could just analyse the activity data from her participants.... we both naively assumed that they knew how to analyse the data they were routinely collecting. But when the student asked them for a bit of help with analysing the data, the response was, you are as much of an expert as we are... [Liz, HCI Researcher]

7.4.2 Physical capability

No factors were found about participants’ physical capability in relation to evaluation.

Overall, industry professionals appeared to be unmotivated to evaluate effectiveness. As well as major motivational barriers, industry professionals were also felt to face barriers in their psychological ability (e.g skills) to evaluate effectiveness, and were constrained by both their physical and social environments. Importantly, it was felt that there may be an opportunity to reduce the amount of time effectiveness evaluations take in the form of automation. The next section describes rapid research designs which have been

proposed to reduce the time taken to evaluate effectiveness, and mostly (although not exclusively) focuses on the perceptions and experiences of researchers and data scientists.

7.5 Motivation to use rapid research designs

7.5.1 Reflective motivation

7.5.1.1 Perceived value of rapid research designs

Those interviewees who were aware of rapid research designs appeared to be interested in using them. One product designer (Mark, the only professional in industry who was aware of the designs and was not a data scientist) spoke about "trying" these "really useful" research designs when there is the "opportunity". While an HCI researcher who did not currently evaluate effectiveness, recognised their utility for future evaluations that did examine effectiveness:

> Microrandomisation... I haven't tried it myself but.... it might, indeed, become more important and, yeah, if you need statistically significant results this would be something that I would look at once it becomes relevant for me... it was beyond my questions. [Fred, HCI Researcher]

Interviewees reflected on the features and capabilities of specific rapid research designs that were valuable. One health researcher felt factorial designs such as MOST and SMART would be "very useful" because apps have "many components" (Joseph, Health Researcher). Joseph went on to describe how when designing a physical activity app there had been "a lot of decisions to make" such as which "behaviour change techniques" to incorporate, and rapid research designs may facilitate more evidence-based decisions. Maria, who had previously conducted a MOST trial, explained that using MOST enabled her research group to understand not only the "individual effects" of components but also their "interactive effects" [Maria, Health Researcher], and that this was useful information particularly for "optimising" interventions.

In addition to MOST and SMART, another health researcher, Tom, described the usefulness of CEEBIT to continuously evaluate and improve the quality of apps that had already been released or "rolled out". Both health and HCI researchers

also considered the characteristics and advantages of using microrandomised trials and single case design or N-of-1 approaches, including their ability to understand 'exactly what works for who' [Michael, HCI Researcher]. Fred, an HCI researcher noted that microrandomisation can investigate app effectiveness for people with rare health conditions and specific health characteristics and "make sure that every dimension [condition and characteristic] is sufficiently represented in your sample".

One health researcher, Tom, felt that "we don't do enough within-subjects designs" generally, including single case design ("N-of-1") approaches that "collect baseline data". He felt that there was the opportunity to do "keep doing quality improvement" with single case designs and "more interactive time series work" by "stagger[ing]" the introduction of different "version[s]" of the app to participants. Another health researcher, Chris, felt that there were parallels between single case design approaches and artificial intelligence methods which collect large sets of data on an individual, and that integrating these could help understand why interventions work for a particular individual.

Maria, a health researcher, highlighted how she had chosen between rapid research designs (which have different characteristics), based on the type of app that was being evaluated:

> There were... adaptive trials, the SMART ones and... N-of-1, but that wasn't what we were going for. So the intervention was always going to be a sort of personalised but not continually changing, the only thing that changed was the feedback given which was updated as they gave us more information. But it didn't adapt to them... it was more constant and constant intervention content as it were. [Maria, Health Researcher]

Thus, trial designs such as SMART may be particularly useful for apps that adapt over time. One health researcher felt that because rapid research designs can assess effectiveness of apps quickly and efficiently, and industry professionals work quickly, rapid designs would be "an easy sell" to industry:

> They were pretty eager to learn this stuff and mostly…they're an easy sell for these alternative designs because you tell them that, yes, I know you only have a 12 to 18-month horizon for your development to

> launch and I'm going to slow it down with a five year RCT, they kind of roll their eyes and look at you like you're nuts. (Tom, Health Researcher).

Interviewees also felt that people required more information about the benefits of these designs, in order to be motivated to use them. They highlighted the need to conduct studies to provide evidence that rapid research designs are efficient and beneficial:

> I think if people...if somebody would have picked up the gauntlet and gone and ran a trial similar to one that's been run in a classical design and rerun the same as that trial within adaptive design and showed that, you know, you could get basically the same results in, I don't know, tenth of the time or something like that then maybe more people will be open to use it... showing that you could significantly save the amount of effort... [Daniel, data science/industry professional]

Overall, being aware of the value of rapid research designs appeared to motivate academic researchers, in particular, to use them.

7.5.2 Automatic motivation

No factors were found that related to automatic motivation for using rapid research designs.

7.6 Opportunity to use rapid research designs

7.6.1 Physical opportunity

7.6.1.1 Time-based factors

Despite their name, researchers perceived rapid research designs to be time consuming. Joseph recalled how he had discussed using SMART designs with colleagues but it was decided that it would not be feasible to organise or "set up" within the time available to complete the project:

> If I may be honest with you about something... I came across SMART designs. So the literature coming out from the US... I was, like, overly excited about it, and I went to my [colleagues], and I said, this is what we could try to do. And yeah, we would never be able to set up a study like that in time, to fit the [project] schedule. So, that

> idea was soon abandoned. So I didn't learn anything about SMART designs, or other more appropriate designs that could be used to test apps [Joseph, Health Researcher]

Another health researcher with experience of using rapid research designs suggested that the MOST approach did indeed take time to implement. Maria described how setting up the experimental conditions for MOST had involved developing several versions of an app to facilitate different intervention components and multiple "control" app versions. Maria also recounted that administrative "checks" were required to ensure users would be correctly assigned to different app versions, which she described as "time consuming". Maria recalled that she did not anticipate the length of time these tasks would take:

> there were sort of just the logistical challenges of developing and designing five intervention modules and the control for each of those. And then checking that... the randomisation was correct and that eligible users were correctly randomised... So that was whilst the trial itself was quicker because you could run them all in parallel, there was still a lot of checks that needed to be done which was something that I hadn't thought about before doing it. It was an incredibly dull and time-consuming task of just downloading the app multiple times.... [Maria, Health Researcher]

Crucially however, it was the preparation for the MOST trial that was time-consuming, and Maria pointed out that the "the trial itself" was "quicker" and a "more rapid way of finding out" which intervention components "work" than RCTs. She explained that using "traditional" RCTs to understand which components were effective would require researchers to run several trials, one after the other, of apps containing one component. In contrast, MOST designs ran these "in parallel" at the same time. She therefore felt that MOST provided "great return" on the time initially invested:

> you have to put in a lot more work, by now saying right we're developing these five intervention models and like setting it all up, but you get such great return [Maria, Health Researcher]

Maria concluded that whether researchers should invest time in planning and conducting a MOST trial depended on their "long term goals". If researchers had "no intention to develop it further" (i.e. optimise and improve the app beyond

finding out whether it was effective) then by conducting a MOST trial "you are just making life very difficult for yourself".

Another health researcher, Tom, described a time-related challenge he experienced when conducting "rapid RCTs" (as described in chapter 6, section 6.3.2). Tom recalled that even when using these designs, some health behaviours, such as smoking, required longer trial durations to accommodate measurement of longer-term outcomes:

> [we] typically did relatively short outcomes... the only problem, in smoking we had to at least do six months in order to be able to have something that would be legitimate literature [Tom health researcher]

Importantly long-term data collection periods were needed to produce "legitimate" findings that were worthy of publication. Tom further noted that otherwise, trial procedures were deliberately shortened ("as short as we could possibly make it").

Overall, different time-related challenges were reported for different rapid research designs.

7.6.2 Social opportunity

7.6.2.1 Funding approval and publication acceptance

A major barrier that was reported by academic researchers and industry data scientists to using rapid research designs was the attitudes and motivations of funders. For example, health researchers felt that many funding organisations would only pay for studies that employed RCTs. Tom had experienced this in the past:

> I would have a hard time to tell you the truth, and maybe it's changing a little bit now, but when I was doing this research it was often the case that I couldn't get a grant approved unless I was a doing two arm RCT... Even though there were cases where there was more appropriate designs that could be used. [Tom, Health Researcher]

Chris also recognised this focus on the RCT and described funders as "conservative":

> it's extremely hard to get anything else but a randomized control trial funded. Funders are very conservative and they like their randomized control trials, and as soon as you come up with something else they will not fund that ... they will not fund your project... It's very, very tough to get money from these guys, very low success rates, and so they only fund the best of the best basically. And, we'll always end up with randomized control trials... any other design doesn't really fly for them [Chris, Health Researcher]

Chris implied that RCTs were regarded as "the best of the best", and felt that a research team would be reluctant to submit an application for an mHealth project employing a rapid research design, as "there's always the potential that it's unacceptable".

HCI researchers reported different challenges in applying for health funding, including the iterative nature of the development and evaluation work involved in mHealth. Liz noted that the need for continuously improving app design and the need to conduct several evaluations was something health funders did not "acknowledge":

> I think, you know, most of the health funding, like NIHR in the UK, or NIH, I guess, in the US, tends to focus a lot on the RCT... and the big numbers. And fails to acknowledge that actually you need to personalise it, you need to iterate on the design, and evaluate the design multiple times, before you can get to do a meaningful trial. And that... the technology has moved on before you've got the results of the trial [Liz, HCI Researcher]

Liz therefore indicates that an RCT would not provide "meaningful" results without this initial iterative design work because the app would not be optimised before the trial. Liz further added that RCTs were "so expensive", emphasising the point that it is not worthwhile conducting RCTs until design work is completed. Another HCI researcher [Michael] similarly faced challenges in receiving funding for projects involving app design work. Interestingly, however, he felt that funders actually welcomed the methodological rigour brought about by rapid research designs including single case designs in the context of iterative

work that is otherwise unstructured and unpredictable ("they get really positive reception").

Participants perceived this conservative attitude to be in place not only in relation to funding, but also in being published. Most strikingly, an industry data scientist reported that in his research company, staff were familiar with and used rapid research designs "internally", yet actively chose not to use them when publishing in academia. This is because they felt the work would not be accepted by "the scientific community":

> To us it was very natural to actually use these kinds of designs.... the problem is that when you try to sell, as it were, the paper to the academic community we felt that it would be very hard for them to buy something like that and so we resorted to something simpler which would be an easier sell.... they're not that acceptable, in my feeling at least, to the scientific community which is unfortunate because...most of my work is on internet data... there is always that concept of adaptive experimentation. [Daniel, Industry Data Scientist].

Daniel further described how "a range" of decisions for designing both the app and the trial were "due to this kind of thought process" where they "decided to make it simple because we thought that if we added more layers of complication it would be very hard to get it accepted" (for example, restricting the number of SMS messages sent and the timing of these).

A widely held belief across interviewees was that rejection would be due to the fact that rapid research designs were "unusual" [Daniel, Industry Data Scientist] and not "traditional" [Maria, Health Researcher] or "typical" [Tom, Health Researcher] of the grant applications and publications that reviewers would normally receive:

> The more...it's not specifically the design that's the problem it's the fact that it's another unusual thing that you would use in the trial and the more these unusual factors come in the harder it is to get it accepted. [Daniel, Industry Data Scientist]

Therefore, it was perceived that that reviewers of grant proposals and academic manuscripts are motivated by established traditions, and participants viewed this to be an important barrier to using rapid research designs. Interviewees felt

that that it was necessary to change funders' attitudes to rapid research designs, and to promote funding decision-making to be based less on accepted norms and more on the appropriateness of a research design:

> It's sort of, what would help people make that move to a broader range of designs and evaluation approaches and getting those funded.... getting reviewers on grants to think more broadly about which designs are most appropriate in these digital interventions and to not penalise people who aren't doing the traditional approach [Tom, Health Researcher]

Liz felt that current funding systems "lack...joined uppedness" in that each only funded specific types of research (either RCTs or "basic" lab-based, non-applied research) neither of which were felt to be applicable to app development work:

> EPSRC fund certain kinds of research, NIHR, and MRC, and Wellcome, and whoever, funds different kinds of research. And I think there is a lack of joined upped-ness.... You know, I've experienced very specific examples of that, where you just kind of go, okay, we know this is the thing that needs to be done, will NIHR fund this – no, because it's not close enough to the RCT. Will EPSRC fund it – no, because it's beyond the basic research. Who else will fund it are...you know, and it just kind of falls down through that. [Liz, HCI Researcher]

Overall, Tom felt that rapid research designs "make so much sense" and are often more appropriate than RCTs for mobile health technologies but that it currently "takes more work to convince a journal reviewer" of this. He felt that "it takes a while I think for people to make that adjustment". As such, efforts to change the thinking and behaviours of funders would need to be substantial and on-going.

7.6.2.2 Colleagues' experience of using rapid research designs

Colleagues who interviewees worked with day-to-day appeared to influence their own use of rapid research designs, both positively and negatively. One HCI researcher, Liz, described how she had learnt a lot of what she knew about rapid designs through working with people from other disciplines who already used them:

> Yes, so I learned about MOST and SMART, from jointly supervising a PhD student, with (Clayton, Feehan et al. 2015), at [UniversityX]... So

> I've learned an enormous amount from working with, particularly (Clayton, Feehan et al. 2015), but also several other health services researchers. So, all, I think all of my PhD students, at the moment, bar one... have their other supervisor from a health faculty. [Liz, HCI Researcher]

Maria, a health researcher who had herself employed a rapid research design, described how this experience had led to other members of her research group being "keen" to use rapid research designs on other projects:

> Within the research group they were very aware of it. I think Researcher1 had previously worked with… somebody in the sort of research group of [someone who developed the rapid research design] and potentially it was just [them].... And so researchers in the team were very aware of that and keen to use this new sort of method of developing and optimising interventions. And so sort of from the outset that was a kind of guiding principle as it were.... And I think I was lucky that researchers were very keen to try something, not quite you know, try something new, that this was a new approach doesn't matter, but that it was the right one for us. [Maria, Health Researcher]

However, colleagues were also perceived to be a barrier to using rapid designs. One health researcher felt that he conducted an RCT because he worked in an institute that typically used "formal" RCTs to evaluate pharmacological interventions:

> So the team there, their background, is very much a clinical trial focus type of work.... they've done lots of clinical trials in terms of medication, and pharmacology type of things. They would be hired to do that type of, very formal randomised control trial, typically two arms, you know, placebo... that's one of the reasons, I think, why we ended up going with the randomised control trial to test these apps. Because, traditionally, that's what the department has done, that's the gold standard, and that's what, you know, it certainly had an impact on our decision [Joseph, Health Researcher]

In addition to the pressure to conduct RCTs, Joseph noted that the lack of colleagues using rapid research designs meant that he could not receive help in conducting them because there was a lack of "expertise" around him.

7.7 Capability in relation to using rapid research designs

7.7.1 Psychological capability

7.7.1.1 Awareness of rapid research designs

Most researchers had heard of at least one type of rapid research design (i.e. CEEBIT, MOST, SMART, SCD/N-of-1, Microrandomised trials). Researchers also perceived that other researchers were becoming "more aware" of these designs [Fred, HCI Researcher]. Some interviewees did not know about the rapid research designs in detail, while others could describe their characteristics.

Both health and HCI researchers admitted that they had not always been aware of rapid research designs and previously "had no idea of their existence" [Joseph, Health Researcher]. Michael, an HCI researcher, told of one occasion when his research group had done a "classical randomised control trial" because "back then" they were not "aware of that kind of set up" (i.e. rapid designs), even though they would have been appropriate for use in place of an RCT:

> well interestingly if at the time when we wrote the grant proposal we would have known about single case designs that we would have done that.... It is absolutely one of the kind of context where we could have easily done single case designs. [Michael, HCI Researcher]

One major factor facilitating the spread of awareness of rapid research designs was academics finding out about them by attending talks at conferences. Michael also noted that conference social media activity also allowed exposure to these research designs both among attendees and more widely:

> I think it's because of the digital health conferences in London that I follow, based on Twitter and on ResearchGate, quite a few of the UniversityX crowd, right, and people who are sort of loosely connected to that [Michael, HCI Researcher]

Published literature (i.e. methodology papers, HCI editorials, textbooks) appeared to be another key facilitator for spreading awareness of rapid research designs. Researchers from both Health and HCI described how they had first come across particular designs in the literature, and had then read up on them in detail.

The industry professionals, and particularly CEOs, interviewed were less familiar with rapid research designs. One CEO noted that he was "not too much into exploring, I have…I more try and inspire the teams to do that" [Steve, CEO], while another CEO who was involved in effectiveness evaluations noted that it was the health researchers he collaborated with who "deal" with the types of experimental design used [Pat, CEO]. Another industry professional, despite having an academic background and being familiar with experimentation, was not aware of rapid research designs as he does not "keep an eye on that specifically" [James, CEO].

7.7.1.2 Experimental Set up

All health researchers expressed difficulties in conceptualising and implementing rapid research designs. Joseph described how differences in experimental procedures from the traditional designs he was more familiar with led to uncertainty in decision making around the setup of the study:

> for someone with my background, it was a bit scary... in terms of, well, the study has to be set up differently, as to what I was more aware of... what the typical study is... I mean, N-of-1 studies, N-of-1 randomisation studies, I don't know how to do those things... And then, we could have done more complicated things, like factorial designs... participants can be exposed to so many different treatments, and how do you assess, and how do you decide which treatments, or components of the intervention people get exposed to? [Joseph, Health Researcher]

Joseph went on to explain that people in his research unit did not know how to use rapid designs. Because time was limited and there was a need to complete the work, this resulted in the team continuing to use the designs they felt most comfortable using:

> There was no expertise within the unit to do it, and because we needed to get something done instead, we just, you know, we did what the expertise was, what our resources were [Joseph, Health Researcher]"

Joseph's account highlights the relationship between different types of barriers. Feeling that he did not have the time, and that others did not have the

"expertise" to support him in using the design, highlights connections between capability and opportunity (both physical and social opportunity, respectively).

7.7.1.3 Statistical analysis

Some interviewees reported not feeling confident in their own ability to analyse the results from rapid research designs. Joseph, a health researcher, regarded himself as having "poor" statistical skills in general. Nevertheless, he felt capable of analysing RCT data, which he associated with "basic" statistics. He perceived the analysis for rapid research designs to be considerably more complex and that they would require specific experience that he did not have:

> And another thing is how you analyse. Like, I don't have, I have a poor background in statistics. So I have learned some of it, I do know the basics, everyone can, you know, learn to more or less analyse RCT data. This analysis, if you do go to SMART designs, and since your participants actually do kind of change group, and not just the factorial... I don't have experience with analysing those data, so I kind of ran away from them. [Joseph, Health Researcher].

Joseph felt he needed help from expert statisticians before committing to using rapid research designs, but time constraints meant that this was not always practical and in the past had prevented his use of potentially valuable designs:

> I tried to approach, statisticians in the university, but we were so pushed in terms of time, that by the time we got to meet with them, we had already made our ethics application, and so there was a design decision already. So when we finally got to meet with the statisticians, it was like, already trying to figure out how we could analyse the data, and so we didn't have any input on different designs that we could have gone with in terms of answering what our research questions could be. [Joseph, Health Researcher]

This suggests a need to understand how data would be analysed before deciding to use rapid research designs, a role for statisticians in discussing alternatives to the RCT, and the overall need to establish interdisciplinary collaborations early in the planning stage of mHealth projects. In addition to the statistical analysis itself, one industry professional highlighted additional challenges requiring early input from statisticians, for example, choosing adequate sample sizes for MOST and SMART factorial trials.

> How do we absolutely ensure we've got [an appropriate] sample size, when we're running the test?... I would like to get a little bit more confidence in before rolling it out, right, running it... I'm sure we'll be wrong, but I would like to be less wrong. [Mark, Product Designer]

Like many other interviewees, Mark ultimately felt he needed to feel more confident and knowledgeable before attempting to use rapid research designs.

7.7.1.4 The need for education and training, learning from others, and tools:

Participants felt there was a need for education not only on the range of rapid research designs and methods available ("What's needed is a better understanding of the range of different techniques that both HCI and health services research can offer" [Liz, HCI Researcher]), but that was also focussed on practical issues, including when and when not to "cut corners" [Liz, HCI Researcher], and how to use rapid research designs in industry contexts (i.e. integrate into a staggered product release, or 'structure the workflow' with industry professionals) [Mark, Product Designer]. Interviewees also expressed a need for statistical training, and some felt that practical software tools facilitating the use of rapid research designs could be an extremely useful support to supplement education and training:

> ...there are some parts of evaluation that you need to tackle by means of education. So, by educating people that N-of-1 study is useful and valid, whereas there are other parts of evaluation that you need to just address by having the right tools. So, where do you draw the line, is the hard question.... how do I solve this? Do I solve this by educating the POs [product owners and designers] and telling them that things don't trend towards significance, they're either significant or not significant, or is it something that you solve by actual tooling so that the tool doesn't show P values, it just says, this is not significant? Or your minimum time of this experiment hasn't been crossed so no results available yet, kind of interaction. So, yeah, in an ideal world we would all be statistical experts but in practice we need to try to internally educate people in data literacy as well as provide them with tools to facilitate this stuff. [Aaron, industry data scientist]

7.7.2 Physical capability

No factors were found that related to physical capability for using rapid research designs.

7.8 Discussion

This chapter presents an in-depth analysis of the barriers and facilitating factors in evaluating behaviour change apps and wearables, and using rapid research designs. Several barriers for industry professionals in evaluating effectiveness were associated with their main motivation to make a profit (prioritising engagement evaluations over effectiveness, perceived lack of time, money and staff to conduct experiments, and limited opportunities due to financial decision makers). Further barriers included risks, and perceived lack of capability to conduct effectiveness evaluations. Together, these barriers and concerns exceeded the potential facilitators of evaluating effectiveness, including the potential for improved company reputation, a possible competitive advantage, and doing "social good". Participants felt that automating effectiveness evaluations could reduce the length of time evaluations require, but automation processes would need to be "bullet proof".

In relation to using rapid research designs, researchers and industry data scientists perceived these to be valuable (i.e. were motivated to use them) and reported factors which improved their awareness. These facilitators, however, were hindered by perceptions and experiences surrounding the length of time needed for preparation and setting up of rapid research designs, limited acceptance by funding bodies and perceived lack of capability. Academic colleagues were both facilitators and barriers depending on their experiences with rapid research designs, and researchers felt that more education and support for the use of rapid research designs was needed.

Industry professionals' ultimate motive to increase profit is not surprising, as this drives commercial businesses. Beyond a lack of motivation to assess their product's effectiveness, successfully demonstrating effectiveness and achieving positive results (and receiving favourable media coverage or doing "good") was not considered to be financially worthwhile. Although benefits to industry in evaluating effectiveness for a range of computerised health systems have been proposed, this was an early opinion piece (Henderson et al. 1999). Empirical work and financial analyses are needed to understand whether and how evaluating effectiveness impacts the finances of a health app/wearables

company, and to explore ways to maximise profit. Results should be actively disseminated to industry professionals who may otherwise remain sceptical of any benefits of evaluating effectiveness.

Industry professionals conceived of two risks in evaluating effectiveness. The first of these is the risk of an app being found to be ineffective. However, failure in relation to effectiveness could be viewed as an opportunity for app redesign. Companies (particularly startups with limited resources) are familiar with risk and the need to 'fail fast' to quickly learn how to improve their product (Giardino et al. 2014) (e.g. rapidly redesigning an app in ways that increase retention). Increasing the agility of effectiveness evaluations to enable "failing fast" may enable industry professionals to quickly learn how to optimize effectiveness and attenuate any risks or negative financial impact of ineffective products. A second risk related to liability and regulation. Physical activity is an example of a behaviour that can have medical implications for particular patient groups, for example in those with diabetes (Eng and Lee 2013), which may fuel uncertainties surrounding regulation.

HCI researchers' mixed views on their role in evaluating effectiveness reflect a wider debate on this topic. Klasjna et al (Klasnja et al. 2011) argue that rather than assessing the overall effectiveness of behaviour change technologies, HCI researchers should focus on whether individual intervention components achieve the precursors of effectiveness (e.g. promote self-efficacy). Others have put forward contrasting views: HCI can offer insights into long-term effectiveness (Smith 2012, Dunton et al. 2014); HCI researchers should evaluate effectiveness (using RCTs) but only when they are appropriate (Cresswell et al. 2017); and that assessing effectiveness increases HCI researchers' focus on the ultimate goal to build an effective product (Stawarz and Cox 2015). Overall, while there is consensus on the role of HCI expertise and methods in evaluating engagement and acceptability (Murray et al. 2016, Michie et al. 2017), there is not yet consensus on their involvement in studies focussed on evaluating effectiveness.

Researchers were interested in assessing effectiveness and engagement in the context in which an intervention is deployed: HCI researchers in the broad context of use (i.e. "in the wild" research (Rogers and Marshall 2017) and health

researchers in exploring methods for developing, disseminating and optimizing technologies with the final setting in mind. While some researchers feel health researchers do not typically acknowledge the need for ongoing development, which is a key principle of HCI (Blandford et al. 2018); the interviews suggest some recognition of this. Industry data scientists were also interested in real world data problems (such as cleaning and sparse data), which is a typical aspect of data scientists' day-to-day work (Provost and Fawcett 2013). These similarities between disciplines are opportunities for "synergy" and may improve the success of collaborations (Pagliari 2007).

Similar barriers relating to opportunity and capability were identified across evaluating effectiveness and using rapid research designs. In addition to industry professionals' financial concerns being a major barrier towards evaluating effectiveness, financial concerns also reduced interviewees' opportunity to use rapid research designs. Grant committees were believed to be more likely to fund traditional RCTs than rapid research designs. Just as commissioners and investors fund industry professionals that can demonstrate a return on investment (via engagement metrics such as downloads), Ioannidis et al note academic funders and committees are also "eager to ensure that they get a good return on their investments; inadequate research diminishes the fiscal investment that they have made". (Ioannidis et al. 2014, p.13) The interviews suggest that one way that research funders believe they can get 'value for money' and reduce the risk of research being inadequate is to fund only RCTs. In both industry and academia, it would be therefore be useful to change the attitudes of the finance providers.

Interviewees felt that industry professionals generally do not have enough time to accommodate lengthy effectiveness evaluations, and rapid research designs were considered time-consuming to use. The latter is particularly surprising given that these have been proposed to improve the speed and efficiency of research (Kumar et al. 2013, Riley et al. 2013). The interviews suggested that it was specifically setting up these designs and learning how to use them which took time. Interviewees' perceptions align with an existing view (Whittaker et al. 2012) that the time required by MOST to choose which features to control and isolate may reduce its superiority over RCTs.

In addition to overlapping opportunity factors, perceived capability in planning and statistically analysing experiments were barriers to both evaluating effectiveness and using rapid research designs. It may be the case that people involved in evaluating mobile health technologies do not know the requirements of rapid research designs. Preliminary research seeking to explain the limited use of SCDs in clinical settings has suggested that SCD requirements may not be fully understood (Kravitz et al. 2009).

The factors identified as barriers to both evaluating effectiveness and using rapid research designs could inform solutions that simultaneously target these behaviours (i.e. the evaluation of effectiveness using rapid research designs). Automation, which was considered to improve the speed of effectiveness evaluations, should also be explored for rapid research designs given they may be time consuming to use. Interviewees conceptualized potential features such as real-time data analysis and long-term data collection. Further research is needed into which aspects of an evaluation to automate: creativity and human input is required not only for designing app features, but also constructing appropriate research designs (Shadish et al. 2002). To guide the use of automated rapid research designs, a framework may be useful. This may be welcomed by academic researchers motivated to use rapid research designs, however further work would be needed to understand how to integrate this approach in ways that are “bullet proof” and do not negatively impact industry professionals’ current practices.

7.9 Conclusion

This chapter explored barriers and facilitators to evaluating effectiveness, and using rapid research designs. Motivational factors differed between these two activities, however, some opportunity and capability factors that appeared to be associated with both evaluating effectiveness and using rapid research designs to do so. Understanding these similarities may enable solutions to be formed that encourage researchers, data scientists, and industry professionals to assess effectiveness using rapid research designs.

In chapter 6, interviewees discussed the use of app store approaches to facilitate pragmatic RCTs, and this chapter presented interviewees' beliefs surrounding the "automation" of evaluations as a possible opportunity to improve the speed and efficiency of evaluations. Chapters 8 and 9 explore an approach that automates an effectiveness evaluation through embedding a particular type of rapid research design (single case designs) within an app store release of physical activity apps.

Chapter 8 A Framework for Operationalising Single case designs for physical activity apps Distributed via App Stores (The OSDAS Framework)

8.1 Introduction

The previous chapter explored barriers and facilitators for evaluating apps and wearables targeting physical activity and other health behaviours, and in using rapid research designs. Researchers and industry professionals reported that evaluations in general, and evaluations using rapid research designs, were time-consuming. However, the automation of study procedures was identified as an opportunity to speed up evaluations. One way to automate a trial is to programme an app to collect necessary data and systematically execute study procedures, and to distribute it via an app store. This approach can enable many real-world users to: download the app; participate in the trial; undergo study procedures and provide data, without requiring direct contact with the researcher. In addition to saving time and increasing efficiency, the automated nature of the trial can: (i) increase the generalisability of any effectiveness results to real-world settings as the study is conducted within these settings; (ii) provide a way to assess the effectiveness of behaviour change apps distributed via app stores; and (iii) potentially enable the continuous assessment of effectiveness over time.

Chapter 6 revealed interviewees' perspectives on the app store approach and chapter 7 highlighted researchers' concerns over their capability to conduct rapid research designs. There is no guidance or frameworks on using rapid research designs to evaluate apps distributed via app stores (i.e. app store apps). Although a small number of app store studies have used RCTs (BinDhim et al. 2014, Volkova et al. 2016), and some rapid research designs (e.g. MOST (Crane et al. 2018), whether automating single case designs (SCDs) in this way would facilitate scientifically rigorous app store studies remains unexplored. SCDs may be particularly useful for HCI researchers (e.g. in targeting users) and other disciplines in mobile health research. This chapter presents the development of a framework that supports researchers in using SCDs to assess

apps (specifically those targeting physical activity) distributed via app stores. The first version of this framework (V1) is presented in this chapter. Following this, chapter 9 reports the refinement and testing of the framework through its application to the design and deployment of a specific physical activity app. The final framework (V2) is provided the end of chapter 9.

8.2 Consideration of suitable SCDs for the OSDAS Framework

There are different types of SCD that vary in their experimental procedures; specifically, how and when the intervention is introduced to the participant and/or withdrawn. The types of SCD which are accepted by guidelines as being able to demonstrate a causal relationship include: reversal/withdrawal (i.e. the intervention is introduced and then withdrawn); multiple baseline (i.e. the intervention is introduced at different times to different participants); changing criterion (i.e. intervention goals gradually increase); and alternating treatments (different interventions are rapidly introduced and withdrawn, i.e. "switched") (Smith 2012, Tate et al. 2013). "Mixed" or "combined" designs are also acceptable, whereby features of different types of SCD are drawn together in one study (Smith 2012).

Choosing a type of SCD to use can depend on: (i) the target behaviour under study; (ii) researchers' aims for the study; (iii) the nature of the intervention; and (iv) the setting in which the intervention is delivered (Smith 2012, Tate et al. 2013, McDonald et al. 2017). Researchers may aim to understand effectiveness across different users (as opposed to a single individual): here, a multiple baseline design would be suitable. The research question may also include assessing the effects of specific intervention components (i.e. app features): both multiple baseline and reversal designs are suitable for such analysis (Dallery and Raiff 2014), either by introducing components in isolation ("drop in") or removing a single feature ("drop out") and observing their effects (Horner et al. 2005). The nature of the intervention itself may aim to increase or 'shape' behaviours gradually, which would be amenable to changing criterion designs (Hartman et al. 2016).

Using SCDs to evaluate mobile apps generally (i.e. not specifically app store apps) is still in its infancy; however, the few studies using this approach have used a variety of SCD types (changing criterion (Kurti and Dallery 2013), multiple baseline (Rabbi et al. 2015), withdrawal (Daskalova et al. 2016). Dallery et al (Dallery et al. 2013) suggests behaviour change app-based interventions can support a range of SCDs. However, these studies all incorporated visits to the research lab, contact with a researcher, detailed study instructions and monetary incentives. The app store setting introduces additional challenges that may restrict which types of SCD can be used. App-based behaviour change interventions themselves often suffer low levels of engagement (Yardley et al. 2016), which can make it difficult to collect sufficient data. Low engagement and difficulty obtaining data is likely to be amplified in trials run entirely remotely, without supportive factors such as contact with the researcher, monetary incentives and detailed study information to manage user expectations (Henze and Boll 2010, Eysenbach et al 2011).

Adequate user engagement is important for SCDs, as all types require regular measurement (Horner et al. 2005). Yet, different types of SCD may promote positive or negative user experiences, and unacceptable procedures may negatively effect engagement. Table 5 outlines the strengths and weaknesses of SCDs in supporting causal inferences (adapted from Dallery et al. 2013), and their strengths and weaknesses in promoting positive user experience in app store settings.

Withdrawal-based designs (e.g. reversal, alternating or combined single designs with a withdrawal period) are particularly robust in assessing the causal effects of an intervention because they provide multiple opportunities to assess effectiveness for an individual (i.e. each time the intervention is introduced and withdrawn). Randomised N-of-1 RCTs, a specific type of withdrawal-based SCD, are considered the gold standard for understanding the effectiveness of an intervention for an individual (Guyatt et al. 2000, Shamseer et al. 2015). Yet, engagement is likely to be particularly challenging for withdrawal-based SCDs. During a remote trial with little opportunity for researchers to explain study procedures and manage expectations, it is not difficult to imagine how intervention withdrawal would lead to a confusing user experience (whereby a

user finds that an app feature, or the entire app, is suddenly no longer available). This could perhaps be perceived as a bug or problem with the app's functionality: i.e. whether the app functions or as anticipated and features remain in the app (i.e. the app' "stability"), which has been considered to be a key factor in HCI (Blandford et al. 2018) and an important concern for users in downloading app store apps (Ferreira et al, 2012).

Type of SCD	**Experimental procedures**	**Strengths and weaknesses in supporting causal inference (i.e. experimental rigour)**	**Strengths and weaknesses in supporting positive user experience (UX) in an app store setting**
Reversal (e.g. ABAB)	Baseline collected, intervention is implemented, and then intervention is removed	Strengths: Strongly supports causal inference through within-subject replication; clear demonstration of an intervention effect in one subject. Weaknesses: Not applicable if behaviour is irreversible. Requires removal/withdrawal of intervention.	Strengths: none Weaknesses: Withdrawal of intervention or intervention features may disrupt UX experience. Users may feel confused without a researcher present to provide an explanation.
Multiple baseline	Typically, participants begin a baseline phase at the same time (i.e. the trial is run concurrently). The treatment is introduced in a staggered fashion across participants.	Strengths: Enables between-subject replication and comparison (i.e. provide indications of whether the intervention is effective for multiple participants). Treatment does not have to be withdrawn. Useful in situations where "learning effects" occur. Weaknesses: Does not typically enable within-subject replication. Users may have to endure long baseline phases without any intervention.	Strengths: Withdrawal of the app or app features is not required. Weakness: App store users may be unlikely to endure a long baseline phase
Changing criterion	Following a baseline phase, intervention goals are implemented in a step-wise manner.	Strengths: Enables within-subject replication (goals act as different levels of the independent variable). Useful when gradual change in behaviour is desirable. Treatment does not have to	Strengths: Withdrawal of the app or app features is not required. Disruption to UX may be minimal for goal-setting apps. Weakness: This design

	Goals become progressively more challenging as they are met.	be withdrawn. Weaknesses: Must have continuous outcomes.	requires long-term user engagement.
Alternating Treatment	Two or more treatments are rapidly "switched" and compared. Treatments can either be compared to a baseline phase, or to each other.	Strengths: Allows comparison between treatments. Weaknesses: Requires removal/withdrawal of intervention while another is introduced.	Strengths: Some participants may find it engaging that app content continuously changes. Weaknesses: Rapid switching between app features and components may disrupt UX experience. Users may feel confused without a researcher present to provide an explanation.
Mixed (or combined) design	Elements of any treatment can be combined	Strengths: Elements from other designs that improve experimental inference and rigour can be selected and combined in ways that overcome individual design weaknesses. Allows for more flexible, individually tailored designs. Weakness: Any negative impacts of above designs could have additive effects.	Strengths: Combining multiple baseline designs and changing criterion designs enables both within-subject and between-subject comparison, and provides a more robust design for a real world app store setting. Weaknesses: users may be unlikely to endure a long baseline phase, and engage in the app over time.

Table 5: **Causal inference and user experience considerations when choosing an SCD for an app store-based trial.**
Columns 1-3 (from left to right) are adapted from Dallery et al. 2013

Multiple baseline designs require long baseline periods for some users (to allow other users to be introduced to the intervention). Rabbi et al, for example, employed baseline and control phases for several weeks with only basic features and without 'active ingredients' (Rabbi et al. 2015). Baseline phases in SCDs may also consist of no intervention (i.e. measurement only periods, Kravitz 2016). Participants downloading an app from the app store may expect particular features immediately and discontinue use. Using a multiple baseline approach

will require designing highly engaging baseline phases, or alternatively, taking advantage of novel features on some phones that do not require users to endure a baseline phase, by using data collected by the phone before the app was installed.

Changing criterion designs necessitate long-term intervention use in order for participants to be exposed to different criterion levels. Kurti and Dallery (2013) required users to complete many consecutive '5 day experimental blocks': engagement may not be sustained for this long on an app store app. Changing criterion would be particularly suitable for apps which incorporate goal-setting features, and are expected to support long-term use.

A mixed design, which combines multiple baseline and changing criterion approaches, may be most suitable for app store deployment. Neither multiple baseline nor changing criterion designs require withdrawing or "switching" interventions or intervention features. Furthermore, incorporating features from both multiple baseline and changing criterion designs enables both between-subjects and within-subjects comparison, which strengthens the ability to make causal claims (Kratochwill et al. 2010, Kratochwill 2014). Multiple baselines can be used to understand effectiveness for multiple participants and individual differences, while several changes in criterion goals provides multiple opportunities to understand if the app is effective for an individual. It is expected that combining these two forms of SCD will provide a robust design for app store deployments.

8.3 Framework overview

The OSDAS Framework is intended to support researchers in Operationalising Single case designs for physical activity apps Distributed via App Stores.[13] Operationalising is defined in this thesis as embedding or implementing experimental requirements (i.e. the criteria and quality indicators needed in order to demonstrate causality), within an app store app. Thus, operationalizing SCD requirements involves programming an app (and arranging an app launch) to

[13] The OSDAS Framework described in this thesis is intended for use by researchers as opposed to industry professionals. Study 2 (presented in chapter 7) indicated that industry professionals experienced specific barriers towards evaluating effectiveness in general. Further research required to tailor OSDAS to industry professionals is discussed in chapter 10.

support SCD criteria and quality indicators. As shown in Figure 4, the 3-stage framework is intended for use after the main components of a physical activity app have been designed and developed, and before analysing whether the app is effective in changing behaviour.

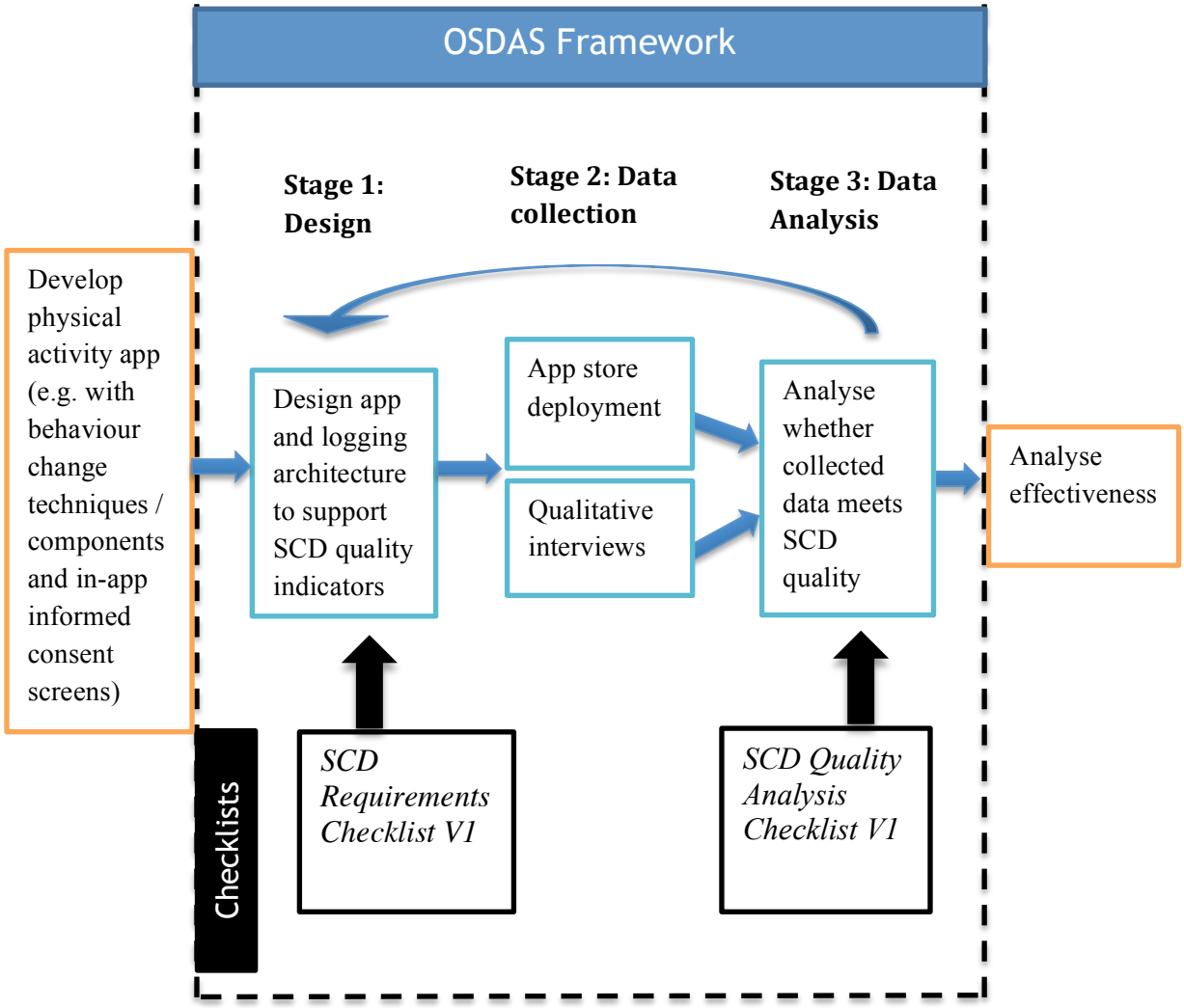


Figure 4: The OSDAS Framework (V1).
Black dotted lines indicate the boundaries of the framework. The orange boxes convey design and analysis activities that would be required before and after the OSDAS Framework is applied. The checklists developed to support researchers in stages 1 and 3 of the framework are version 1 (and are refined in chapter 8). In stage 3, data collected from an app store deployment is analysed (visually and statistically) to verify whether data is of sufficient quality to support effectiveness conclusions.

The development and design of the physical activity app, although outside of the scope of the framework, may influence how researchers go on to operationalise SCD requirements. As discussed above, the OSDAS Framework is based on one

type of SCD (i.e. multiple baseline changing criterion design), which is particularly suitable for apps containing goal-setting components. Hekler and colleagues (Hekler et al. 2016) describe an agile process for designing and developing apps to include behaviour change techniques, which involves understanding user needs and prototyping (i.e. rapid iterative development) to arrive to at a functioning app. In addition to behaviour change techniques, it is assumed the app will include in-app information and consent screens. This process of "implementing ethics" in an app store app is described elsewhere (Rooksby et al. 2016).

The OSDAS Framework itself consists of three stages. Stage 1 ("Design") involves designing the physical activity app to incorporate SCD requirements. To support this stage, the "SCD Requirements Checklist" was developed, which outlines the quality indicators that were collated across established SCD checklists and presents these in relation to five key criteria (dependent variable, independent variable, baseline, internal validity, external validity, social validity). Specifically, the SCD Requirements Checklist guides the design of specific features the app should incorporate, and, the design of the data logging architecture (i.e. programming the app to log particular user interactions and sensor data).

Stage 2 of the framework is the data collection phase. This follows the 'hybrid' approach proposed by Morrison et al (2012), which consists of deploying the app on an app store and collecting device-generated user logs (as specified in stage 1), and conducting user interviews to gather qualitative data. In the OSDAS Framework, interviews are recommended for addressing particular quality indicators (i.e. those relating to 'social validity').

Stage 3 of the framework involves analysing the extent to which the data collected from Stage 2 meets SCD requirements (i.e. assessing whether the research design was implemented as planned and whether the data can be used to assess app effectiveness in improving physical activity). To support stage 3 of the OSDAS Framework, the "SCD Quality Analysis Checklist" was developed, which draws on SCD requirements to outline key questions that should be asked of the data collected.

Analysing effectiveness and effect sizes are beyond the scope of the OSDAS Framework; rather, the framework focuses on ensuring that any claims regarding an app store app's effectiveness, which are made using an automated, fully-remote single case research design, are likely to be scientifically valid.

If the data meet SCD requirements then researchers should proceed with effectiveness analysis. If the data do not meet SCD requirements, then researchers may return to (stage 1) to alter the design of the app and logging architecture to better meet requirements. Otherwise, the researcher may proceed with effectiveness analysis and include only users who met criteria.

8.4 Framework stage 1: operationalising SCD criteria within an app store deployment

The first stage of the OSDAS Framework involves designing and adapting an existing physical app to incorporate SCD requirements using design methods such as prototyping (Wilson and Rosenberg 1988, Maguire 2001). Designs will be based on existing app components and features (e.g. how and when behaviour change components are introduced to users) and also involve developing new components (such as a baseline phase) and the logging architecture.

Version 1 (V1) of the SCD Requirements Checklist is presented in Table 6, which outlines SCD criteria and quality indicators collated from existing checklists (as described in section 4.6.1.1). These included: the What Works Clearinghouse Standards (Kratochwill et al. 2013); the Single Case Experimental Design (SCED) criteria (Tate et al. 2008); Risk of Bias in N-of-1 Trials (ROBINT) (Tate et al. 2013), the APA Division 16 Task Force on Evidence-Based Interventions in School Psychology (Kratochwill et al 2003), and those described by Horner et al. (2005), Dallery et al. (2013) and Klein et al. (2017). While some quality indicators are based on a study's research design (e.g. whether a baseline phase is used), others must be evidenced by analysing and reporting collected data (e.g. whether a sufficient number of data points were collected during the baseline phase) (Kratochwill et al. 2010). Thus, the SCD Requirements Checklist outlines quality indicators to include within the design of the app, and, the extent to which collection of trial data is required to provide evidence of study quality.

Criteria Group	QI no.	Quality indicators (QI)	Data collection required?*
Dependent variable (DV)	1.1	DV is described with operational precision and is measured with a procedure that generates a quantifiable index	No
	1.2	DV is repeatedly measured over time, at regular intervals (i.e. with equal increments between each measurement).	No
	1.3	Sufficient number of data points are collected within baseline and intervention phases (minimum of three/five)	Yes
	1.4	Data is collected or referenced on the validity and reliability of dependent variable measurement	No
	1.5	In the case of remote data capture, the identity of the source of the DV is authenticated or validated	Yes
Independent variable (IV)	2.1	IV is described with replicable precision	No
	2.2	IV is systematically manipulated and under control of experimenter, and is continuously implemented over time (Changing criterion)	No
	2.3	If multiple treatments or intervention components are examined, each component is introduced separately	No
	2.4	Fidelity (delivery and receipt of intervention) is measured	Yes
Baseline	3.1	Baseline phase that provides repeated measurement of the DV is included	No
	3.2	Baseline conditions are described with replicable precision	No
	3.3	Baseline lengths vary across participants (Multiple baseline)	No
	3.4	Baselines are independent (Multiple baseline)	No

	3.5	Baseline establishes a stable pattern of responding	Yes
Internal validity	4.1	The design provides at least three replications of experimental effect at three different points in time	Yes
	4.2	The design controls for common threats to internal validity	Yes
	4.3	Participants are randomised to different experimental sequences (e.g. baseline and intervention phase lengths)	No
	4.4	Participants and assessors are blinded to the phase of the intervention	No
External validity	5.1	Design supports replication of the experiment across participants and settings	Yes
	5.2	Participants and critical features of the setting are described with sufficient detail (e.g., age, gender, health condition, therapeutic setting).	Yes
	5.3	The process for selecting participants is described with replicable precision.	No
	5.4	Procedures for ensuring generalisability of results over time are implemented or described	No
Social validity	6.1	The dependent variable (target behaviour) is socially important	Yes
	6.2	The intervention and experimental procedures are acceptable	Yes
	6.3	The independent variable in implemented in a way that is practical and cost effective, by typical intervention agents, in typical physical and social contexts.	No

Table 6: Version 1 of the SCD Requirements Checklist (V1)
*Is trial data required to provide evidence of study quality?

8.4.1 Dependent variable

The first set of SCD quality indicators (QI) (Table 6) relates to the dependent variable (DV): the behaviour that the intervention targets. In reporting a study, researchers should describe how the DV is operationalized and measured with enough detail to allow other researchers to replicate the study, and SCDs should only be used when the dependent variable can be quantified (QI 1.1). The DV

should be measured repeatedly over time to "sample" the behaviour (Tate et al. 2008, Tate, Perdices et al. 2013), and the sampling intervals should be regular (i.e. equal time between each measurement) and predetermined before the study commences (e.g. once every 24-hour period, as opposed to randomly) (Christ 2007) (QI 1.2). There must be enough data points in each phase to allow researchers to demonstrate experimental effects. All guidelines and checklists state that a minimum of three data points per phase is acceptable, however some propose that five is preferable for a high-quality design (Kratochwill et al. 2010, Tate et al. 2013) (QI 1.3). Data should also be collected, or referenced (Kratochwill 2014), on the validity and reliability of the measurement instrument or on inter-rater reliability if human observers are used)(Horner, Carr et al. 2005) (QI 1.4). Dallery et al (2013) also include a relevant quality indicator for capturing data remotely via smartphone; the authors suggest that researchers "authenticate" the data collected to ensure they are collected from the intended individual using, for example, biometric fingerprinting (QI 1.5).

8.4.2 Independent variable

The independent variable(s) (IV) should be described with enough detail to allow other researchers to replicate the study (Horner et al. 2005) (QI 2.1). For SCDs, the IV is the experimenter's manipulation of when and how the intervention is introduced to the participant. These represent different experimental conditions or 'phases', including a baseline phase (whereby participants are measured before receiving any intervention) and intervention or treatment phases. In a changing criterion design the IV should be continuously manipulated (by increasing the criterion or goal over time) to create sub-phases (QI 2.2). If an aim of the study is to examine the effectiveness of different intervention components (i.e. to examine its "active ingredients") (Horner et al. 2005, Ward-Horner and Sturmey 2010, Dallery and Raiff 2014), then these components must be introduced separately (i.e. in isolation) within a new phase (known as the "cardinal rule") (Tate et al. 2008) (QI 2.3). Some checklists outline requirements to measure the "fidelity" of the independent variable. While Horner et al (Knoblauch) describe this as the extent to which the intervention has been delivered as planned (e.g. by intervention agents such as teachers), Dallery et al. (2013) draw on wider conceptualisations of fidelity to include both delivery and receipt of the intervention by participants, i.e. participant's exposure and

use of the intervention in real world settings. (Borrelli 2011, Dallery et al. 2013). (QI 2.4)

8.4.3 Baseline phase

An important feature of SCDs is the baseline phase, in which the dependent variable is measured repeatedly over time in advance of the participant receiving any intervention (QI 3.1). How this baseline phase is implemented should be reported in with sufficient detail for it to be replicated by other researchers (QI 3.2).

Two baseline criteria were found that were specific to multiple baseline designs: multiple baselines should vary in length (QI 3.3) and be independent of each other (Kazdin and Kopel 1975) (i.e. data is collected from independent individuals) (QI 3.4). If the study design requires participants to begin the intervention at the same time (i.e. concurrently), then varying baseline length should be achieved through 'staggering' the introduction of the intervention over time across participants. If participants instead begin the experiment and enter the baseline phase at different times (i.e. non-concurrently), baseline lengths should be predetermined and assigned to participants (Watson and Workman 1981, Christ 2007).

This baseline phase is the control or "counterfactual" condition of SCDs, which provides an estimation of how the dependent variable measure would continue if the intervention were not introduced (Shadish et al. 2002). Data from this baseline phase is therefore used within analyses to establish a pattern of responding, which is then projected into the intervention phase and compared with the actual response observed (Kratochwill et al. 2010). To be a high quality baseline, the data should demonstrate high stability (i.e. low variability in the dependent variable) and no trends (i.e. systematic increases or decreases in the dependent variable), especially in the direction of the desired intervention effect (Kratochwill et al. 2010, Kratochwill 2014). A baseline that already indicates a participant was improving before the intervention was introduced does not provide compelling evidence that the intervention was responsible for the improvement. (QI 3.5).

8.4.4 Experimental control/internal validity

In SCDs, an experimental effect is shown if the dependent variable changes when, and only when, the intervention is introduced. Therefore, internal validity is enhanced by: (i) features of the research design that can reveal experimental effects; and (ii) analysis of outcome data (Kratochwill, Hitchcock et al. 2010). The central means of enhancing internal validity in SCDs is through repeated observations of experimental effects, known as "replications" (Horner et al. 2005, Dallery et al. 2013). Study designs should facilitate at least three opportunities for any effects to be replicated at different points in time (Kratochwill et al. 2010). Different types of SCD vary in how replication is facilitated. Multiple baseline designs across participants should include at least three participants (Kratochwill et al. 2010).[14] Changing criterion designs should support at least three criteria (e.g. exposure to at least three goals after a baseline phase) with two criterion changes (e.g. whereby the researcher increases the goal for a single participant at least two different points in time (Hartmann and Hall 1976, Kazdin 2011, Kratochwill et al. 2010). (QI 4.1)

Research designs should also address 'threats' to internal validity (i.e. rival explanations for any effects observed). For multiple baseline designs, an important threat to address is the 'history' effect (Christ 2007)): the risk that a change in the dependent variable could be attributed to a confounding, external event that affected all individuals simultaneously, as opposed to the intervention (Shadish et al. 2002).[15] Therefore, multiple baseline studies should provide "verification periods" across participants, whereby the baseline phase of one participant overlaps with the intervention phase of another (Carr 2005) (see Figure 5). A multiple baseline verification period is typically achieved through staggering the intervention across time (i.e. introducing some participants to the intervention while others remain in the baseline). Design features for changing

[14] Multiple baseline designs can alternatively include replications across three *behaviours* for a single participant (i.e. within-subjects), however the OSDAS Framework supports a between-subjects multiple baseline design. The exact number of replications that should be included within an SCD has been debated: earlier studies prior to guidelines and checklists suggest a minimum of two participants (e.g. Baer et al 1968, Kazdin & Kopel 1975)

[15] Carr (2005) notes that some 'non-concurrent' multiple baseline designs (Watson & Workman, 1981), where participants begin the experiment at different times, are less able to demonstrate control against the 'history' effect because they do not typically enable these verification phases. As later described, the thesis examines whether verification periods occurred even though participants begin the study at different times.

criterion designs should particularly address the “maturation” threat (Hartmann and Hall 1976), which is the risk that participants would have improved even without exposure to the intervention (Shadish et al. 2002). Verification for changing criterion designs is facilitated through varying the length of sub-phases, and the size or magnitude of the change in criterion, to demonstrate that the participant is not simply improving their behaviour at a steady rate. (QI 4.2).

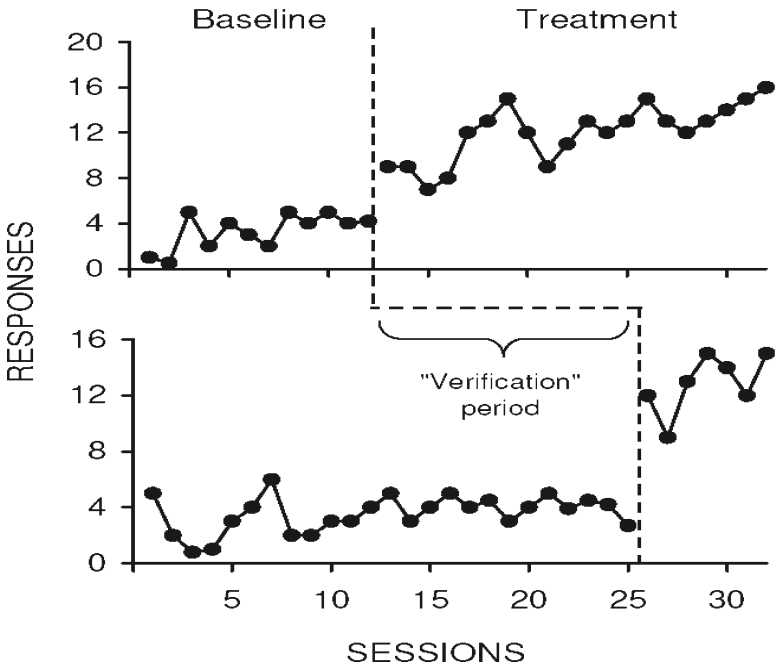


Figure 5: A verification period for a multiple baseline design with hypothetical data
Taken from Carr et al., 2005

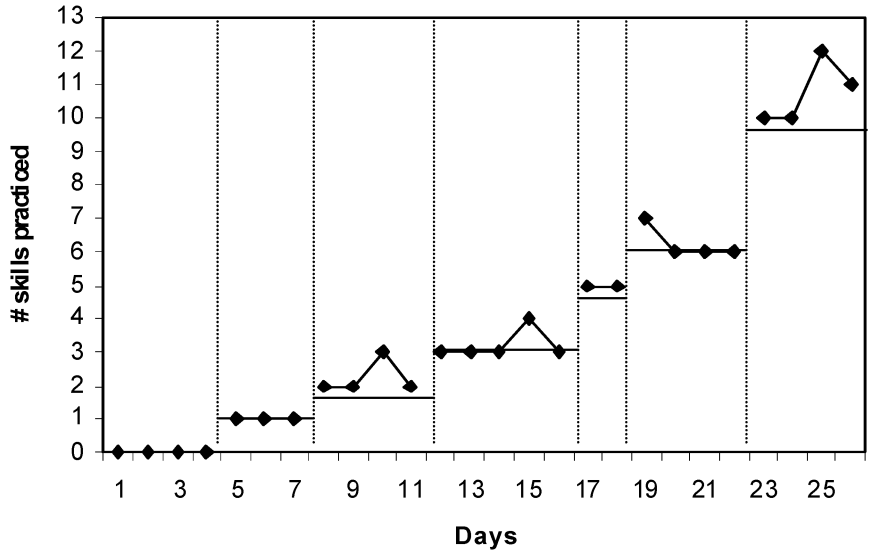


Figure 6: **A changing criterion design where increasing levels of goals are met by a participant with hypothetical data**
Taken from Rizvi & Nock, 2008.

Two checklists included randomization as a quality indicator of internal validity in SCDs (Tate et al. 2013, Kratochwill 2014). Others emphasise that randomization is not required (Kratochwill, Hitchcock et al. 2010), as replication is the central means of demonstrating internal validity (Dallery et al. 2013). SCED, an earlier version of the ROBINT scale, does not include randomisation and notes that this may wrongly 'penalise' studies for which it is inappropriate (Tate et al. 2008, p. 396). Although randomistion is included in the APA 16 checklist (Kratochwill 2014), the authors note that randomization can be difficult in field trials and is not sufficient to overcome validity threats. Yet, it is generally recognised that randomization can be used to strengthen internal validity (Kratochwill et al. 2010) and can be embedded within multiple baseline designs by randomly allocating participants to different baseline lengths or starting times (Watson and Workman 1981). Randomisation is therefore included in V1 of the OSDAS SCD Requirements Checklist.

Only one checklist, the ROBINT scale (Tate et al. 2013) includes 'blinding' of participants or assessors. For a study to be of high quality, participants should not know whether they are in the baseline control or experimental intervention phases and neither should the individual analysing results. The authors note that blinding the participants can be difficult (although possible with emerging technologies) and thus collapse blinding participants and blinding assessors within one quality indicator. The need for blinding was not included within other guidelines and checklists.

8.4.5 External validity

Checklists and guidelines varied in whether they included criteria specifically addressing external validity. The WWC guidelines advocate the use of separate studies to understand external validity, conducted by different 'research teams' at 'different institutions' (Kratochwill et al. 2010). Dallery et al. (2013) do not explicitly incorporate external validity when outlining the main criteria, although they describe how studies can be replicated across different participants and settings to increase generality. Horner et al (2005) include one external validity quality indicator relating to replication: to replicate the study "across participants, settings, or materials" in order to enhance generalisability of findings (QI 5.1). This can also ultimately support analysis about who the

intervention works/does not work for. Horner et al (2005) outline separate criteria relating to "description of participants and setting" in order to allow researchers to replicate a study, and the APA 16 Task Force (Kratochwill 2014) describes similar criteria as "descriptive and supplemental". However, the criteria described by Horner et al (2005) and APA 16 Task Force provide a means of directly measuring external validity. As such, they were collapsed under "external validity" and included: describing participants (e.g. age, gender and other demographics and physical setting in detail (such as therapeutic setting or classroom) (QI 5.2); describing the process by which participants are selected for inclusion in the trial (including if and how participants are excluded based on unsatisfactory baseline data (Christ 2007) (QI 5.3); and describing procedures for ensuring generalisability over time (maintaining the change in behaviour, through e.g. booster sessions) (Kratochwill 2014) (QI 5.4).

8.4.6 Social validity

Checklists and guidelines varied in whether they included criteria relating to social validity. The WWC guidelines, SCED and ROBINT scale did not include any criteria relating to social validity. Dallery et al (2013) propose that social validity can be considered in terms of: (i) the social importance of the dependent variable (i.e. target behaviour) (QI 6.1); (ii) the acceptability of the intervention and experimental procedures (QI 6.2); and (iii) the social importance of the magnitude of change in the DV resulting from the intervention (QI 6.3). Dallery et al., (2013) also discuss the need for social validity to address data privacy concerns, and Horner et al (2005) describe how intervention acceptability can be understood through exploring participants' intent to continue using the intervention, and whether they perceive the intervention to be effective. Although not categorised as "social validity", the APA Task Force include one ("supplementary") criterion on acceptability of the intervention and procedures, and in relation to the magnitude of change in the outcome, require changes in the DV to be 'clinically significant' (Kratochwill 2014). A recent systematic review of social validity within SCDs (Snodgrass et al. 2018) and previous methodological literature (Wolf 1978) also categorise social validity in these three dimensions (i.e. social importance of the DV; acceptability of intervention/procedures and social importance of results). Horner includes further criteria describing how social validity can be improved through the

independent variable being implemented in a way that is 'practical and cost effective', and in "typical physical and social contexts". These criteria were collapsed to one quality indicator (QI 6.4).

Researchers can use the SCD Requirements Checklist to design the app logging architecture: a structured list of all the specific data the app is designed and programmed to collect. Table 9 summarises basic types of log data from which a logging architecture for an app aiming to increase users' physical activity might be based. Most are "user logs" that automatically record users' interaction with the device (delivery of app features is a system log that records particular actions initiated by the device).

Data collected	Explanation
Unique identifier	The logs used to ensure a dataset is associated with one user may vary depending on the app store used. Apple does not permit identification of individuals: only individual app downloads.
User sensor data	Data retrieved from the in-device sensor within the smartphone to record physical activity.
User interaction data	Logs indicating whether users interacted with the app on a given day, whether particular app features were used, and any demographic data collected entered by the user.
Timestamp and timezone data	Date and time for every user interaction with the app, including the day on which the app was downloaded from the app store.
Delivery of app features (system logs)	For example, in a goal-setting app, logs of the goal value(s) generated.

Table 7: Main types of log data to inform the logging architecture of a physical activity app store app

8.5 Framework stage 2: Data collection

Table X summarises the basic types of log data that researchers should programme the app to collect.[16] Quality indicators for social validity can be addressed by collecting qualitative data as opposed to log data. Morrison and colleagues (Morrison et al. 2012) provide considerations for different qualitative methods that can be used alongside an app store release. While the authors ultimately recommend face-to-face interviews with "local" users (i.e. those who can visit the lab, as opposed to "global" app users who may be geographically

[16] Prior to releasing the app to the public on an app store, researchers may wish to conduct internal testing. This involves distributing an app to selected individuals such as research team members via a platform other than an app store (e.g. a sending a link to the app via email) in order to test data collection and the logging architecture.

dispersed), the OSDAS Framework aims to support fully remote trials. Therefore, telephone or Skype interviews may be better suited. Interviewees may be existing app users, or those who download the app specifically for the purposes of the interviews.

Researchers should report the date on which the app was deployed on an app store. This allows analyses in OSDAS stage 3 to account for any external events that took place during the data collection period (e.g. sporting events, holidays), which may influence users' physical activity (i.e. to examine history effects). Researchers should also report if and when an updated version of the physical activity app was released on the app store (e.g. after completing OSDAS stage 3) and record any changes that were made to the data collected. Major changes might require researchers to produce separate data sets for each app store launch, while minor changes may allow data to be amalgamated to form a single dataset, for analysis within OSDAS stage 3.[17]

8.6 Framework stage 3: Analysing data to validate study quality

The quality of SCDs is determined not only by implementing certain research design features discussed above, but also by the quality of the data collected from the study. For example, the collection of three data points per phase may be planned, but not achieved. Parker and Vannest (2012), influential SCD methodologists, describe how this logic differs to group-based designs:

> ... unlike group research, initial specification of a strong [SCD] design does not ensure that the resulting design will be strong. Regardless of the strength of a specified design ''on the shelf'', the obtained data must be visually examined for unexpected patterns... scrutiny of data patterns within and across phases may lead to the conclusion that despite appropriate planning, the resulting design has degraded from strong to weak, and is unable to support a causal inference. [page 256]

Parker and Vannest (2012) explain that even when a research protocol is designed to be "strong" (i.e. designed to enable researchers to be

[17] As proposed by Mohr et al (2015), changes made to the user interface (as opposed to back end changes to the logging architecture) that affect the delivery of intervention components may threaten internal validity and require a separate analysis.

confident in their conclusions surrounding causal inference and intervention effectiveness), assessing data quality permits an understanding of whether the intended research design "survived data collection" (p.58). Examining data quality can also reveal whether any corrective statistical techniques are required (Parker and Vannest 2012). For example, if an increasing trend was found in some participants' baseline data, "detrending" may be applied: this involves using statistical manipulations to remove and account for trends in a data set to ensure the data is stable (see QI 3.5) before any effectiveness analysis. To manage missing data, statistical techniques such as multiple imputation (i.e. replacing missing data with imputed values based on non-missing data) and full information maximum likelihood estimation (i.e. estimating population parameters using non-missing data) may be required (Smith, 2012). Alternatively, participants with unstable or missing data may have to be excluded (Watson and Workman 1981). As systematically excluding users from analysis based on their data threatens the validity of any effectiveness claims (i.e. mortality bias), researchers should report the reasons for exclusion and possible influences on validity (Christ 2007).

To prevent the need to exclude users, researchers may employ formative, on-going assessment (i.e. close observation of data throughout the duration of the trial) to detect data problems as they occur (Ledford, Lane et al. 2018). This enables researchers to actively manipulate the study design and increase study quality before the study ends. Guidelines suggest that on detecting unstable data or trends within baseline phases, for example, researchers should extend the baseline phase until stability is reached (Kratochwill et al. 2010). However, for a remote app store trial that can potentially recruit many participants, on-going formative analysis by the researcher and mid-study design changes tailored to each participant are unlikely to be feasible[18]. Instead, the study is highly reliant on app users performing particular behaviours and the app functioning as intended.

[18] In future, it may be possible for formative analysis to be automated within an app store launch. This would involve incorporating artificial intelligence methods, whereby an agent makes design decisions based on current user data (such as extending the baseline phase on detecting instability).

The following sections now describe analysis and reporting requirements for each of the items within the SCD Quality Analysis Checklist (Table 6). Corresponding quality indicators drawn from the SCD Requirements Checklist are provided in brackets.

8.6.1 Dependent variable: Are there a sufficient number of data points within baseline and intervention phases (QI 1.3)? Can the identity of the source of the DV be authenticated or validated? (QI 1.5)

While an app may be programmed to collect a particular number of data points, evaluators should analyse and report whether the minimum acceptable amount was actually collected during the trial. Data collection in an app store setting relies on the app logging architecture functioning as intended (i.e. transmitting data remotely), as well as users using the app and allowing researchers to access their data. Whether or not these behaviours occurred should be reported.

Criteria group	**Quality indicator (QI) to test**	**Corresponding SCD Requirements QI number**
Dependent variable	Are there a sufficient number of data points within baseline and intervention phases	1.3
Dependent variable	Can the identity of the source of the DV be authenticated or validated	1.5
Independent variable (IV)	Was the intervention delivered and received as intended?	2.4
Baseline	Can baseline data be used to predict patterns of future performance?	3.5
Internal validity	Did the design facilitate at least three replications at three points in time?	4.1
Internal validity	Did the design facilitate at least three replications at three points in time and	4.2

	control common threats to validity?	
External validity	Was the experiment replicated across participants and settings?	5.1
External validity	Can participants and settings be described?	5.2
Social validity	Is the dependent variable socially important?	6.1
Social validity	Are intervention and study procedures acceptable?	6.2

Table 8: Version 1 of the SCD Quality Analysis Checklist (V1)

8.6.2 Independent variable: Was the intervention delivered and received as intended? (QI 2.4)

Although an app may be programmed to execute intervention phases at particular times (e.g. to deliver different intervention components), mixed changing criterion/multiple baselines designs rely on users using the app for a long enough period to experience these phases and be exposed to, or 'receive', the intervention components. The study also relies on participants discovering and using different components (i.e. app features) voluntarily. It is important to explore whether these user behaviours occurred (e.g. through usage logs), as well as to determine the functionality of the app in delivering the intervention (e.g. whether correct goal values were delivered).

8.6.3 Baseline: Can baseline data be used to predict patterns of future performance? (QI 3.5):

Baseline stability will depend on users' target health behaviour patterns. Graphs can be visually examined for 'bounce' in the data, where the DV sharply increases or decreases (Lane and Gast 2014), and stability can also be quantified in numerous ways. Guidelines suggest using standard deviations and/or ranges in the dependent variable (Kratochwill et al. 2010) for a single participant, however these measures cannot usefully quantify stability across several participants. An alternative method involves developing 'stability envelopes'

(Gast and Spriggs 2010) whereby 80% of a users' baseline data must fall within 20% or 25% of the mean or median value of that phase to be considered stable. One method that may be particularly useful for app store-based research is Schoenfroed's "relative stability" (variation around the mean over six days) (Schoenfeld et al. 1956, Costa and Cançado 2012). An advantage of this method is that researchers can set different stability criteria (e.g. 5%, 10%, 20%, 30%, 50%). While 5% variability may be used for high quality lab studies (Schoenfeld et al. 1956), researchers can choose whether and by how much to relax this criteria for real world app store settings. Furthermore, setting lower stability criterions can enable more participants to be included in the trial. However, researchers should carefully consider the degree to which they relax criteria: for example, setting stability criteria to 30% would simultaneously lower confidence in any effectiveness results obtained.

Trends within baseline phases can be assessed visually for individual users, using lines of best fit superimposed onto graphs; and other visual analysis methods have been developed especially for SCDs such as the split-middle technique (White and Haring 1976).

8.6.4 Internal validity: Did the design facilitate at least three replications at three points in time and control common threats to validity (QI 4.1, 4.2)?

An app store trial can potentially recruit many participants worldwide, providing numerous opportunities for replication of the experiment. Studies should report the number of users that consented to participate in the study, and record when in time the user downloaded the app and participated in different phases (i.e. through time-stamp log data). Verification periods are points in time that researchers can verify effectiveness results, by comparing data between users starting an intervention and those in the baseline phase. Whether and how verification periods are implemented will depend on operationalisation for a particular physical activity app. For example, if multiple baseline verification periods rely on user download patterns, whether these verification periods occurred should be explored. Changing criterion designs will require reporting the number of criterions changes that actually occurred, which may rely on the

length of app engagement as well as whether users met their goals (assessed via user logs).

8.6.5 External validity: Was the experiment replicated across participants and settings and can they be described? (5.1, 5.2).

App store trials enhance external validity by taking place within the real-world environment, with apps potentially being used in diverse settings worldwide. It is important to assess and describe (rather than assume) variance in participant characteristics. This relies on users providing this information (e.g. in-app or through a questionnaire), and the proportion who did so should be reported.

8.6.6 Social validity: Is the dependent variable socially important? (QI 6.1), and are intervention and study procedures acceptable? (QI 6.2)

Assessing social validity will require assessing users' own perceptions and experiences using qualitative data analysis methods. Different themes will emerge depending on the intervention and study procedures used.

8.7 Beyond the OSDAS Framework: analysing effectiveness

If results from stage three of the OSDAS Framework indicate that the majority of quality indicators can be demonstrated, then researchers would go on to assess the effectiveness of the behaviour change app. Analysing effectiveness is beyond the scope of the framework, however it would involve the assessment of: (i) whether a 'functional relation' exists between the intervention and the target behaviour (i.e. Is the intervention effective? Does the behaviour change when and only when the app is introduced in at least three different points in time?); and (ii) the calculation of effect sizes (i.e. how effective is the intervention?) (Kratochwill et al. 2010, Ledford et al. 2018).

Assessing effectiveness involves comparing data across participants' baseline and intervention phases. Specific outcomes to assess include: comparing and projecting trend (and stability of that trend) between phases, and differences in

their “level” (mean or median value); immediacy of effect (i.e. whether behaviour appears to change soon after the intervention is introduced, as opposed to display a “lag” in the intervention effects); overlap in the values of the DV in the baseline phase and intervention phases (with a greater degree of overlap indicating little change in behaviour); and consistency of any effects (i.e. whether multiple individuals appear to show a similar pattern of effectiveness) (Fisher et al. 2003, Kratochwill et al. 2010, Kratochwill 2014).

Guidelines are available that demonstrate how visual analyses can be used to assess the above outcomes (Lane and Gast 2014). While some SCD guidelines suggest that visual analyses can be sufficient for a high-quality design if conducted systematically (Tate et al. 2013) (Kratochwill et al. 2010), statistics are required to facilitate effectiveness comparisons across several participants (Manolov et al. 2014) and to detect small effects (Van Gemert-Pijnen et al. 2014). Visual analysis may be conducted with a proportion of users (e.g. selected randomly) to understand some of the data patterns included in the trial before eventually conducting statistical analyses.

Some statistical methods quantify visual analysis into metrics to assess the degree of overlap (Scruggs et al. 1987, Kratochwill et al. 2010, Shadish et al. 2014). Other methods aggregate participant data to provide an estimate of whether the intervention is effective across users, such as randomisation tests (which require the research design to incorporate randomisation) and regression-based techniques such as the Ordinary Least Squares (OLS) or Generalized Least Squares (GLS) (Maggin et al. 2011). Many meta-analyses simply report the mean percentage of non-overlapping data across studies (Schlosser et al. 2008). To understand individual differences and what works for whom, more advanced techniques, such as multilevel modelling (Shadish et al. 2013), enable quantification of effectiveness within- and across-participants.

If data from stage 3 does *not* satisfy the SCD analysis checklist, researchers can either: (i) take only those who provided suitable data forward to effectiveness analysis (i.e. excluding users with unsuitable data); or (ii) discard all data (i.e. go back to the design of the app, further optimise it, relaunch a new app version and collect new data). The latter option increases the likelihood of attrition

bias, but there may become a point where it is no longer efficient to continue to optimise the app.

8.8 Discussion

This chapter reports the development of the three-stage OSDAS Framework to support researchers in operationalizing high quality SCDs for app store apps, in particular those targeting physical activity. Existing guidelines and checklists were found to differ in the quality indicators and criteria they propose studies must address if they are to conduct and report a high-quality SCD. These quality indicators and criteria were collated to form two OSDAS Framework checklists. This discussion section focuses on the types of physical activity apps for which the OSDAS Framework can be applied and used.

The OSDAS Framework focuses on SCDs. These rapid research designs have unique advantages, including a focus on assessing effectiveness for particular individuals (Dallery et al. 2013, Johnston and Johnston 2013). However, as other rapid research designs are available, it is important for researchers to consider whether SCDs are appropriate for the app under study. A good indicator that an SCD is unsuitable is if several quality indicators are found to be difficult to operationalise in stage 1 of the OSDAS Framework. For example, researchers may find is not possible to design and modify the existing physical activity app to incorporate a baseline phase. While guidelines and recommendations outline different types of rapid research design (Kumar et al. 2013, Murray et al. 2016), a practical framework (which perhaps encompasses OSDAS for SCDs) could be useful to enable researchers to efficiently decide which rapid research designs to use, depending on the nature of the app and research question.

The type of SCD on which the OSDAS Framework is based is a ‘mixed’ design that incorporates multiple baseline and changing criterion approaches. The benefits and weaknesses of different types of SCD for apps deployed on the app store were described. Changing criterion designs are particularly suitable for interventions that incorporate goal-setting techniques (Michie et al. 2011), and other behavioural change techniques (BCTs) that promote gradual behaviour change (e.g. shaping). The usefulness of the OSDAS Framework should be explored for physical activity apps that incorporate different BCTs. Indeed,

researchers and intervention developers should select BCTs on their anticipated ability to successfully change behaviour (as opposed to their "evaluability" (Leviton et al. 2010), which would involve restricting the design of the app because it cannot be evaluated for effectiveness). This presents an interesting trade off. It is important not to stifle innovation or prevent researchers from making their apps publicly available via app stores, yet it is also important to be able to evaluate their effectiveness in these real-world settings - the OSDAS Framework provides a means of doing this.

Conclusion

A combined multiple-baseline changing criterion design appears promising for assessing the effectiveness of physical activity apps distributed via app stores. This chapter has outlined the three stages of the OSDAS Framework and associated checklists for designing and evaluating the rigour of SCDs in physical activity app store apps. Considerations were also provided for how findings from the framework can inform effectiveness analysis. To ensure that the SCD criteria outlined in the checklists can be feasibly implemented in the design and development of a physical activity app, the OSDAS Framework should be tested. If necessary, the framework can then be refined and tailored in ways that account for natural constraints on research designs imposed by real world settings (Lyon and Koerner 2016). The next chapter therefore presents empirical testing and a refined version (V2) of the OSDAS Framework.

Chapter 9 Applying the OSDAS Framework – The Case of Quped

9.1 Introduction

The previous chapter presented version 1 (V1) of the OSDAS Framework. This chapter reports empirical research conducted to test and refine the framework. Testing involved addressing the following research questions:

> What are the challenges and trade-offs of operationalising a single case design for a specific physical activity app store app?
>
> What are the challenges and trade-offs using data from an app store deployment to support a single case design study?

Refining the OSDAS framework involved first identifying quality indicators that could not be fully operationalised within the design and deployment of Quped: a sensor-based physical activity app. Then using these findings, the OSDAS Framework was adapted to include two sections: "essential" criteria (i.e. those that could be operationalised within the design and deployment of Quped) and criteria "to be considered" (i.e. those that could not be operationalised). Log data collected from an App Store deployment of Quped and user interviews are used to explore challenges that researchers may encounter in using the OSDAS Framework.

9.2 The Quped App: Overview

9.2.1 Behaviour change techniques

Quped is a mobile app developed for iOS devices (and thus available only to iPhone users) that was released on the Apple App store. Quped was designed to support three main behavioural strategies to improve physical activity (Michie et al. 2009):

> Self-monitoring: The app automatically collects daily step count data and presents this to the user.

Personalised goal setting: The app calculates and presents a personalised weekly goal based upon step history.

Social comparison: The app enables users to compare their steps and goals with other users by age and gender

Screenshots from the app are shown in Figure 7. There are three main views. Firstly, to the left of figure 1 is the “Steps view”. This shows the users’ current step count for that day and the average number of steps they have taken every day that week. A goal for the week is also shown, as physical activity guidelines recommend physical activity on a weekly, as opposed to daily, basis (Haskell et al. 2007). The goal is presented as a daily average rather than minimum threshold and therefore the user can compensate for days with few steps by walking more on other days.

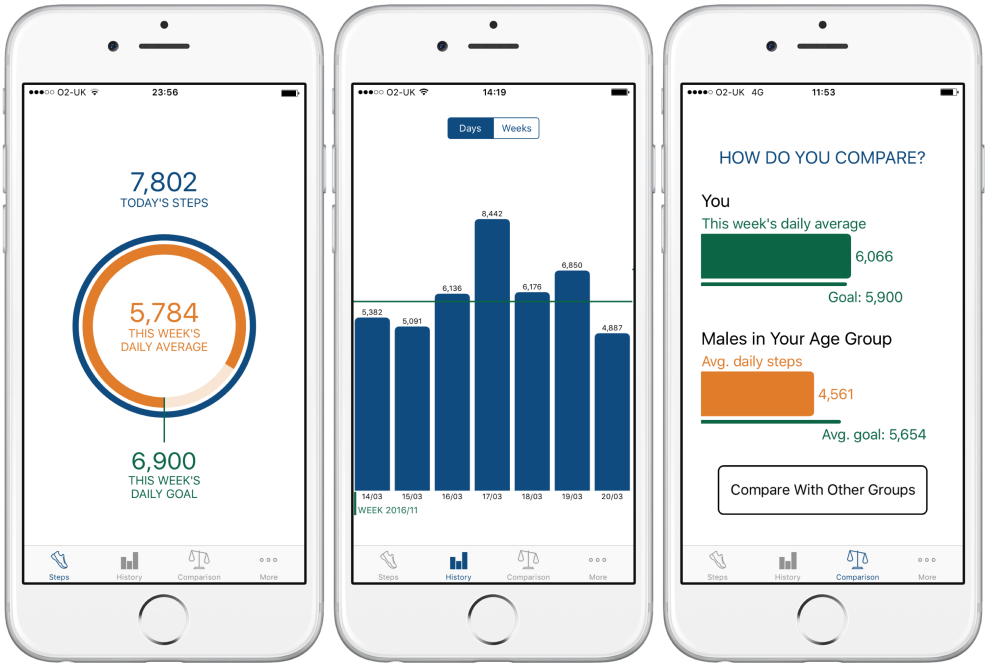


Figure 7: The Quped app
(Left: Steps view. Middle: History view. Right: Social comparison view.)

The personalised goal is automatically generated on the first Monday after installing, and is based upon the averaged value of the user’s previous week of data (which is retrieved upon installation using an iPhone feature, as later described). Thereafter, the value of the goal is determined every Monday morning based upon the average number of steps achieved in the previous 7 days: if users achieved their step goal, they are encouraged to walk an extra

1000 steps above this goal on every day of the following week[19]; if the average step count from the previous week was 95%-99% of their goal, then the daily goal increases by 500 steps, and if user's average was less than 95% of their goal, then the goal for the next week remains the same.

The second main view (middle in Figure 7), is the "History view". In this view, the user can see bar charts of their daily and weekly steps over the period the app has collected data. A line representing the goal at the time the steps were taken is also presented. The third main view (Figure 7) is the "Comparison view". To access this feature, participants must have provided their gender and age. They can then compare their own daily average and goal for the week with the average weekly counts and goals for other users that have also provided their gender and age. Comparisons are grouped by gender and age bracket (organised in increments of 10 years).

9.2.2 Automated research trial features

Quped was designed to support an automated research trial. Figure 8 outlines how the different experimental phases are integrated within the design of Quped. To prevent users from having to endure a baseline phase (as a period with minimal app components and features may not be tolerated in app store settings and result in the user discontinuing use) a feature exclusive to iPhones was employed. This feature retrieves users' step count data from the previous week before the app was installed (this is stored on the in-device sensor: Apple's co-motion processor). Thus, when the user downloads the app, baseline data is retrieved, and simultaneously phase B initiates (whereby self-monitoring and social comparison are available). Phase C initiates the first Monday after installation when the goal-setting feature becomes available. Overall, only phase B and C are actively experienced by users.

[19] This increase of 1,000 steps is also used in an intervention designed by Dallery and Kurti to increase physical activity (and assessed using a changing criterion design). The authors note that the approach aligns with guidelines to increase physical activity gradually and the value of 1,000 was informed by their pilot work (Kurti and Dallery, 2013)

The app was designed to enable the trial to be conducted fully remotely, with minimal researcher involvement and contact. Upon first launch, app screens inform the user that the app is for research purposes, provide information about the study and how the app works, and ask the user if they consent to participate. Data is only collected if a user consents to participate. Users who indicate they are under 18 years of age can continue to use the app features, but no data is transmitted to university servers. Users may also withdraw at any time, by selecting the withdraw option in the "more view". The ethics of the Quped app are further discussed elsewhere (Rooksby et al. 2016).

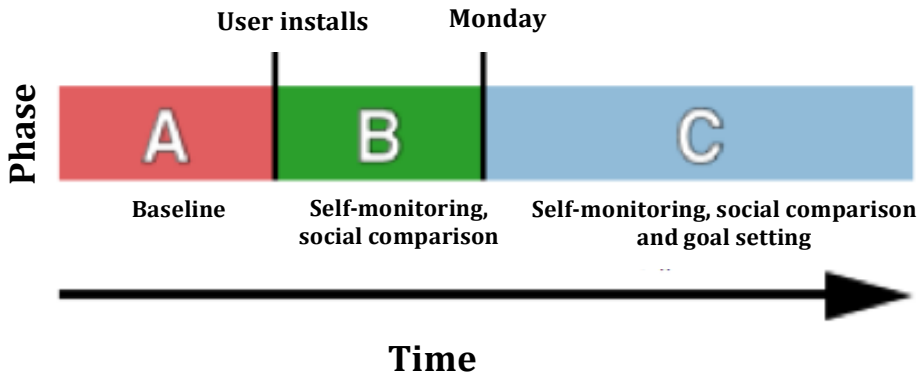


Figure 8: Experimental phases for the App Store trial of Quped

9.3 Framework stage 1: Operationalising a single case design for a physical activity app

As described in chapter 7, a combined multiple baseline and changing criterion design was used in the Quped app. Table 1 outlines the extent to which each OSDAS Framework quality indicator (QI) was operationalised within the design and deployment of Quped. It demonstrates that most (84.0%, 21/25) quality indicators were at least partially operationalised within the app store deployment of Quped: several (64.0%, 16/25) were fully operationalised and some (20.0%, 5/25) were partially operationalised. Only 16.0% (4/25) quality indicators were not operationalised. The following sections describe for each criteria group whether and how quality indicators were operationalised.

Criteria Group	QI no.	Quality indicators (QI)	Operationalised	Reasons for non- or partial operationalisation (if

			in Quped?	applicable)
Dependent variable (DV)	1.1	DV is described with operational precision and is measured with a procedure that generates a quantifiable index	Yes	
	1.2	DV is repeatedly measured over time, at regular intervals (i.e. with equal increments between each measurement).	Yes	
	1.3	Sufficient number of data points are collected within baseline and intervention phases (minimum of three/five)	Yes	
	1.4	Data is collected or referenced on the validity and reliability of dependent variable measurement	Yes	
	1.5	In the case of remote data capture, the identity of the source of the DV is authenticated or validated	No	Nature of the app store: cannot provide participants with equipment that collect physiological 'signatures', app store regulations prohibit using a device ID. Technical feasibility: would require substantial development time. Data privacy concerns: involves storing sensitive and identifiable data
Independent variable (IV)	2.1	IV is described with replicable precision	Yes	
	2.2	IV is systematically manipulated and under control of experimenter, and is continuously implemented over time (Changing criterion)	Partially	Nature of app store: cannot control if/when users download the app. However, automation enables intervention components to be systematically introduced to participants
	2.3	If multiple treatments or intervention components are examined, each component is introduced separately	Yes	
	2.4	Fidelity (delivery and receipt of intervention) is measured	Yes	
Baseline	3.1	Baseline phase that provides repeated measurement of the DV is included	Yes	

	3.2	Baseline conditions are described with replicable precision	Yes	
	3.3	Baseline lengths vary across participants (Multiple baseline)	Partially	Technical feasibility: the step count retrieval technology used did not enable active manipulation of baseline phase. However, baseline lengths can be varied during analyses
	3.4	Baselines are independent (Multiple baseline)	No	Nature of app store and data privacy concerns: app store regulations prohibit using a device ID, as user would be identifiable
	3.5	Baseline establishes a stable pattern of responding	Yes	
Experimental control/internal validity	4.1	The design provides at least three replications of experimental effect at three different points in time	Yes	
	4.2	The design controls for common threats to internal validity	Partially	Nature of the app store: cannot control when users download the app. However, nature of app store also likely to facilitate verification, as users likely to download app at different times. The amount by which the goal changes does vary, but this depends on performance.
	4.3	Participants are randomised to different experimental sequences (e.g. baseline and intervention phase lengths)	No	Technical feasibility: the step count retrieval technology used did not enable manipulation of baseline phase length, and length of phase B varied depending on download date
	4.4	Participants and assessors are blinded to the phase of the intervention	No	Technical feasibility: although users' not aware of baseline phase it is not possible to facilitate user blinding, as app features will change
External validity	5.1	Design supports replication of the experiment across participants and settings	Yes	
	5.2	Participants and critical features of the setting are described with sufficient detail (e.g., age, gender, health condition,	Partially	Nature of app store and data privacy: a detailed description of users requires them to send sensitive and identifiable

		therapeutic setting).		data. However, users can enter gender and age details, and time stamp data identifies whether or not users are based in the UK
	5.3	The process for selecting participants is described with replicable precision.	Partially	Nature of app store: participants self-selected as they choose to download Quped. However, all participants required to own iPhone 5S and above, and would be selected for inclusion in analysis according to data quality
	5.4	Procedures for ensuring generalisability of results over time are implemented or described	Yes	
Social validity	6.1	The dependent variable (target behaviour) is socially important	Yes	
	6.2	The intervention and experimental procedures are acceptable	Yes	
	6.3	The independent variable in implemented in a way that is practical and cost effective, by typical intervention agents, in typical physical and social contexts.	Yes	

Table 9: Single case design criteria and whether they were operationalised in the Quped app

9.4 Dependent variable

(QI 1.1) The first quality indicator for the dependent variable (DV) is that it is described with operational precision and is measured with a procedure that generates a quantifiable index. The DV in the context of the Quped app was user step counts, which were measured automatically using internal sensors embedded within users' own smartphones (Apple's motion coprocessor on iPhone 5S and above). Step count data from the motion coprocessor is automatically stored and accessible via Apple's Health app.

(QI 1.2) The second quality indicator for the DV is that it is repeatedly measured at regular intervals over time (i.e. with equal increments between each measurement). Step values were logged at midnight for each 24-hour period.

(QI 1.3) The third DV quality indicator is that there is a minimum of three/five data points within baseline and intervention phases. The app was programmed to collect six data points for the baseline phase. Intervention phase B was expected to vary between one and seven days depending on when users downloaded Quped (e.g. if downloaded on a Sunday, users would be exposed to only one day of phase B, before phase C initiated the following day), thus, it was accepted that only a subset of users would provide ≥ 3 data points in phase B. Within intervention phase C, a minimum of 5 data points was expected.

(QI 1.4) The fourth DV quality indicator is that data is collected or referenced on the validity and reliability of dependent variable measurement. Validation data was not collected, however, previous studies of the iPhone motion coprocessors compatible with Quped (and that were available during data collection)[20] have been found to accurately assess step counts within lab conditions (except for slow walking speeds (Major and Alford 2016, Duncan et al. 2018). iPhones 6 and above were found to underestimate the number of steps taken within free-living conditions (Duncan et al. 2018)[21]. Duncan et al. (2018) found acceptable *reliability* across iPhones 6S and above at all walking speeds.

(Q1.5) The final DV quality indicator to establish "authenticity" of the data was not operationalised in the Quped study for several reasons relating to the nature of the app store, technical feasibility and also data privacy. Conducting the study remotely over the App Store meant it was not possible to provide participants with any physiological signal equipment as suggested by Dallery and colleagues (Dallery, Cassidy et al. 2013)) or to expect participants to own these already. Substantial development time would have been required to implement any in-app features that could provide physiological readings (or to integrate with 3rd party apps that do so). Not operationalising this quality indicator avoided users having to send sensitive and identifiable data remotely, or the need to store this data on University servers.

[20] These include the "M7" motion coprocessor within the iPhone 5S Major and Alford (2016). and M8, M9 and M10 within iPhones 6, 6S, 6S+, SE, 7 and 7+ (Duncan et al. 2018)

9.5 Independent variable

(QI 2.1) The first quality indicator for the independent variable (IV) is that it is described with replicable precision. A diagram of the experimental phases as defined in the context of the Quped app store trial is shown in Figure 8. The operationalised IV is the onset of the goal-setting component.

(QI 2.2) The second quality indicator is that the IV is systematically manipulated and under control of experimenter, and, as a changing criterion design was employed, this IV should be continuously implemented over time. It was not possible to fully control the introduction of the intervention, as participants' exposure to the entire app depended on if, and when, they downloaded it from the App Store. However, it was possible to systematically control and automate when different app features are introduced. The continuous implementation of the IV is the onset of each new goal.

(QI 2.3) The third IV quality indicator is that intervention components are introduced separately. This involved designing two main intervention phases (as previously described in 9.2). The first (phase B) commences at the time of download, whereby only self-monitoring and social comparison features were available. The app then systematically introduces goal setting the following Monday (phase C).

(QI 2.4) The final IV quality indicator is that fidelity (delivery and receipt of intervention) should be measured. The logging architecture for Quped was designed to measure fidelity: intervention delivery is measured by logging the goal values delivered to users, and receipt is measured by logging whether users have interacted with the app on any given day (and have thus received the self-monitoring and goal setting features automatically shown on the home screen when launching the app). Whether or not participants used the social comparison feature was also logged.

9.6 Baseline

(QI 3.1) The first quality indicator in relation to the baseline phase is that it includes repeated measurement of the DV, and (QI 3.2) the second is that the

baseline is described with replicable precision. The baseline phase (A) for Quped takes place the week before the app is first installed. Using Apple's "Core Motion" feature, step counts for the preceding week are retrieved from the in-device sensor (i.e. motion coprocessor) at the time of download.[22] Internal testing revealed a technical error whereby the first of the seven days often included a step count value of zero, and so the baseline was restricted to the previous six days (which is still above the preferred five data point minimum required).

(QI 3.3) The third baseline quality indicator is that baselines are varied in length across participants. The length of the baseline could be varied between 3 and 6 days "post-hoc" during data analyses by allocating the number of baseline days included for each participant[23]. The length of phase B, which acts as a control period to Phase C, naturally varied between 1 and 7 days depending on date of install.

(QI 3.4) The fourth baseline quality indicator is that baselines are independent. This quality indicator could not be operationalised. App store regulations prohibit the use of unique device identifiers. Unique installation identifiers (UID) were logged for each user; however, if users downloaded the app more than once or on multiple devices this would result in correlated UIDs (i.e. multiple UIDs associated with one individual). Therefore, it cannot be known or tested (without searching for matching data sets within later analyses) whether baselines are independent

(QI 3.5). The final quality indicator in relation to the baseline phase is that it demonstrates a pattern of responding that can be used to predict the pattern of future performance (i.e. that baseline data is stable). Logged step count data

22 This baseline design was chosen over other prototypes in which a user would actively endure a baseline phase. These alternatives included a "dormant" app period (whereby the app simply collects steps without an interface), as has been used within other physical activity studies (Harries et al. 2013). These approaches were considered problematic for the app store setting, whereby users would be unlikely to endure even a short baseline phase with few or zero features (especially when anticipating a fully functioning physical activity app).

23 Specific baseline lengths are not provided by guidelines, however baselines should be long enough to demonstrate stability. Meredith et al. (2011) employ similar variance in baseline lengths (between 2 and 6 days).

from Quped can be used within analysis conducted in stage 3 of the OSDAS Framework to check for stability and trends in the baseline.

9.7 Internal validity

(QI 4.1) SCD research designs must support the ability to determine experimental effects at three different points in time (i.e. facilitate three or more "replications" of experimental effects). Different types of SCD support replication in different ways. For multiple baselines designs, at least three individuals must be introduced to the intervention at different points in time. Launching Quped on the app store enables multiple individuals to download the app and participate in the experiment, and these participants are expected to download the app on different days (i.e. different points in time). This will facilitate replications of any changes in step counts between baseline and intervention phase B. As goal setting may be activated on the same Monday for several users (regardless of whether they downloaded the app on e.g. the Tuesday or the Friday the preceding week), users must download the app across different weeks in order to replicate experimental effects between intervention phases B and C.

To facilitate replication in changing criterion designs, at least three criteria (i.e. with two increases in the goal) must be implemented. In order to be exposed to three step goals (with two increases in step goals), users must remain engaged with the app for three consecutive weeks from the onset of the goal-setting function, and their average step count must be at least 95% of their step goal.

(QI 4.2) The second quality indicator in relation to internal validity is that the research design controls for common threats to internal validity, by providing opportunities to verify any effects found (i.e. any observed changes in behaviour). For multiple baseline designs, verification periods are created through staggering the introduction of the intervention over time. As it is not possible to control when real world users download app store apps, introduction of the intervention could not be actively staggered across participants. However, it verification periods were considered likely to naturally occur within the download patterns across users (i.e. one user will download the app and begin phase B during another users' baseline phase A). It was accepted that

verification periods between phase B and phase C would not occur, as all users currently in phase B would go into phase C simultaneously. However, download patterns may provide verification periods between phase A and phase C, which would still permit examination of any history effects (i.e. through observing whether step counts within one user's baseline phase increased at the time at which intervention phase C was introduced to another).

In changing criterion designs, verification is facilitated through varying the length of time a participant is exposed to each goal and the amount by which the goal changes. To support this, Quped step goals vary in how much they increase per week (either not at all, 500 steps or 1000 steps) and weeks where the goal does not change lengthens the time during which the user is exposed to that goal. However, this variance is necessarily dependent on whether users meet goals.[24]

(QI4.3). The third internal validity quality indicator is that participants are randomised to different experimental sequences. Within multiple baseline designs, this could be different baseline and intervention phase lengths (whereas in a withdrawal design the intervention may be withdrawn at different time points across participants). No randomisation was incorporated into the Quped study. Participants could not be allocated randomly to baseline lengths due to the decision to use step count retrieval technology. All participants received intervention phase C on a Monday, and this meant that the length of the preceding phase B varied depending on the day of the week users downloaded the app.

QI 4.4. The final internal validity quality indicator is that participants and assessors are blinded to the phase of the intervention. This was not operationalised within the Quped study. During the week that baseline data was collected from, users were not aware this was a baseline phase (as it occurred before they downloaded the app). However, users experienced a change in app features between intervention phase B and C. In relation to blinding assessors, researchers in an automated app store trial do not manipulate experimental phases (and are thus unlikely to influence results); phases are implemented by a

pre-determined algorithm. Assessors must analyse specifically baseline data within stage 3 of the OSDAS Framework.

9.8 External validity

(QI 5.1) The first quality indicator in relation to external validity is that the design facilitates replication across different participants (i.e. with varied characteristics) and settings. Incorporating the app store within the research design provided the potential to recruit real-world users with different characteristics and different settings (e.g. downloading and using the app in different locations). (QI 5.2) The second external validity quality indicator is that participants and critical features of the setting are *described* with sufficient detail (e.g., age, gender, health details, therapeutic setting). Participants in the Quped study could be described only with limited detail, as app store regulations prohibit collecting data that will not be useful to the user[25]. The social comparison feature meant that participants would be provided with potentially valuable information (how they compare to others) if they provided gender and age details, however, no further details (such as health information) were collected. To reduce data privacy concerns, participants' locations were not logged. However, each user interaction logged was accompanied by time zone data that could indicate whether or not they were based in the UK when using Quped. (QI 5.3) The third external validity quality indicator is that the process for selecting participants is described with replicable precision. The nature of the app store also hampered the ability to select participants; participants were highly self-selected (downloaded the app on their own accord). The limited ability to collect data on participant characteristics restricted the ability to describe participants "with replicable precision". Generally, inclusion and exclusion criteria were not relevant for Quped as it was not intended for use by any specific population. The only inclusion criteria for selecting participants was that they owned an iPhone 5S and above [26].

[26] Later analysis shows that some participants would perhaps need to be excluded based on the quality of their baseline data. Such criteria would need to be incorporated as inclusion and exclusion criteria in future single case design studies that go on to analyze effectiveness.

(QI 5.4) The final quality indicator of external validity is that procedures for ensuring generalisability of results over time are implemented or described. The nature of the app store enabled the study to run over several months, enhancing the ability to understand effectiveness over time.

9.9 Social validity

The first two social validity quality indicators (QIs 6.1 and 6.2) were explored through analysing qualitative interview data. The resulting themes (outlined in stage 3, Table 12) provide operationalised quality indicators relating to physical activity apps on app stores.

(QI 6.3) The final social validity quality indicator is that the IV is implemented in a way that is practical and cost effective in typical physical and social contexts. Social validity is enhanced by the remote and automated nature of the study: intervention phases are manipulated with minimal impact on users, and users can download and use the app in their own physical and social environments.

9.10 Framework stage 2: Data collection

Having designed the app and accompanying logging architecture to accommodate the above criteria and quality indicators, the Quped app was then released on the Apple App store in February 2016. Log data from the first six months of the app store release were collected for the current study.

During the six-month log data collection period, Quped was updated (in July 2016). The updated version retrieved step data from the users' smartphone in a different way. The in-device sensor itself did not change, rather Quped queried the Apple Health application that retrieved sensor data, as opposed to the sensor itself. This was in response to missing data issues that occurred if users did not open the app for periods longer than a week. The app user interface was also modified to display users' daily distance travelled, as well as their step counts. In the Quped study, data from different app versions were amalgamated for stage 3 of the OSDAS Framework, however researchers going on to analyze the effectiveness of an app would need to consider whether app modifications

would influence embedded BCTs (and thus internal validity of effectiveness conclusions).

In addition to the app store deployment, semi-structured interviews were conducted with Quped users to address quality indicators relating to social validity.

9.10.1 Results

9.10.1.1 Participants

9.10.1.1.1 App store trial

As shown in Figure 9, 222 users[27] downloaded Quped from the Apple App Store. While 65.3% users provided consent to take part in the study (145/222), several (41.4%, 60/145) had to be excluded from analysis because they accessed the app on incompatible devices (i.e. iPads or earlier iPhone models).

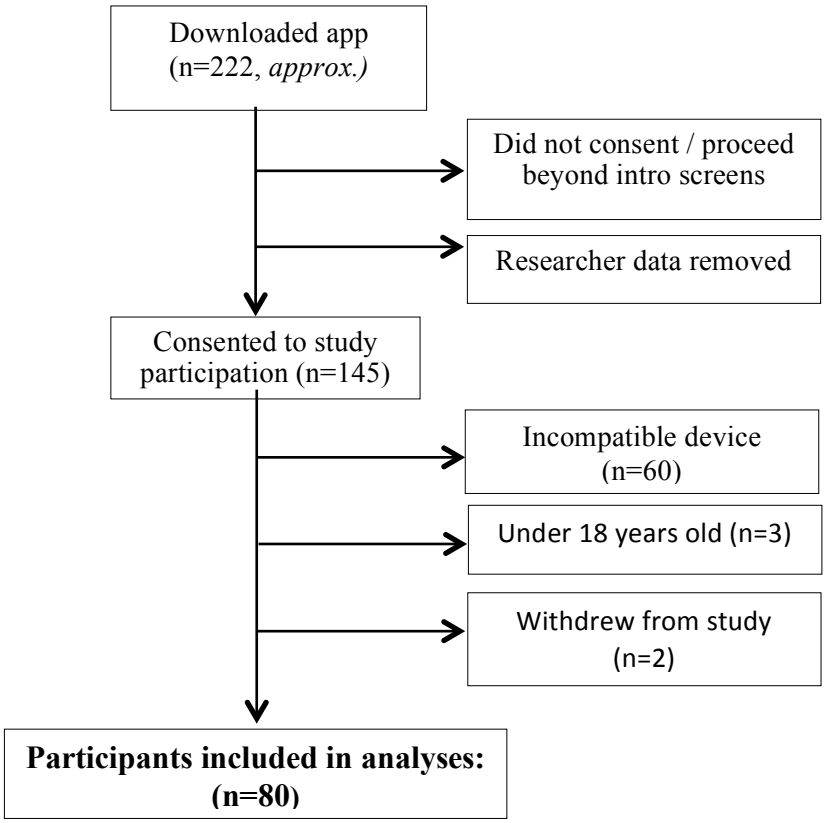


Figure 9: Flow diagram showing the number of users who downloaded Quped in the first six months of the App Store deployment and provided log data for analysis.

9.10.1.1.2 Interviews

[27] Estimated using Google Analytics software

Thirteen of those who downloaded Quped during the first six months of the app store release were interviewed. A further five participants were interviewed after this six-month period to further explore social validity (log data from these participants are not included in analyses). Interviewees included ten males and eight females; most (n=11/18) were 25-60 years, the remainder were 18-24 years.

9.11 Framework stage 3: Data Analysis

While the previous section described the extent to which the design of Quped supported an SCD, framework stage 3 reports whether data from the App Store trial of Quped meet quality indicators in the SCD Quality Analysis Checklist. Quality indicators relating to the dependent variable, independent variable, baseline, internal validity and external validity criteria (QIs 1.1 – 5.4) were analysed using log data collected from the app store deployment. Social validity quality indicators (QIs 6.1 and 6.2) were analysed using qualitative data from interviews.

9.11.1 Dependent variable: are there a sufficient number of data points within baseline and intervention phases (QI 1.5)?

Missing data (i.e. days with step count values of zero) was found in users' baseline (16.3%, 13/80) and intervention phases (42.5%, 34/80). Examples of users' step data with zero values in different intervention phases are provided in Figure 10 and Figure 11. Nevertheless, 86.3% (69/80) provided at least three data points in the baseline phase, and the majority provided the minimum number of data points required in intervention phases: 46 (48/80, 57.5%) contained at least three data points in phase B and 48 (48/80, 60.0%) contained at least five data points during phase C. Overall, 53.8% (43/80) provided data sets with sufficient data points in all three phases. Importantly, data missing in phase C (whether intermittently, or as a block of missing data as shown in Figure 11) creates challenges in using the changing criterion design to assess effectiveness, as this requires exploring data when receiving goal setting components.

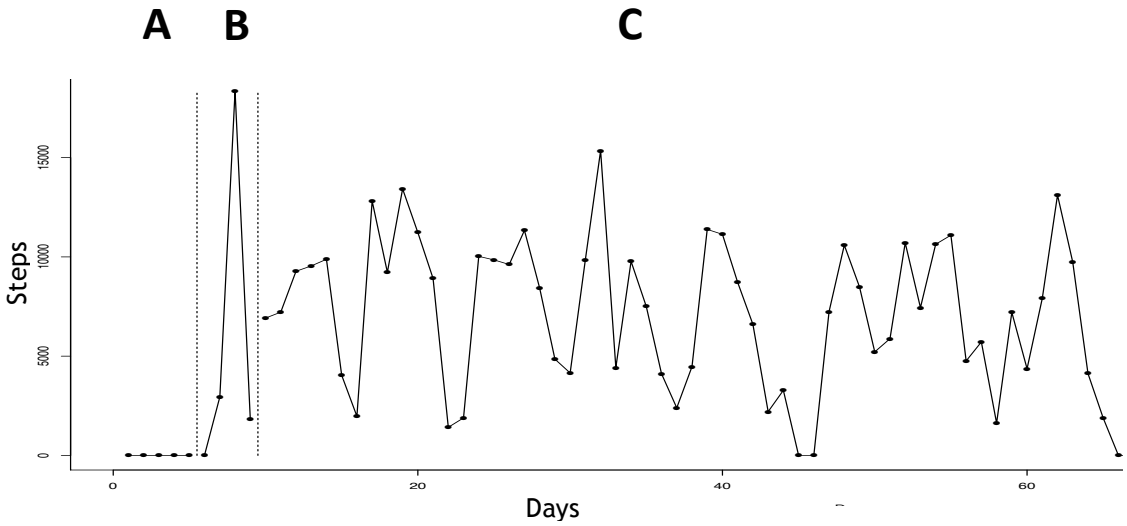


Figure 10: A user's step count data in phases A, B and C (zero values example 1)
Zero values occurred in the baseline phase ("A")

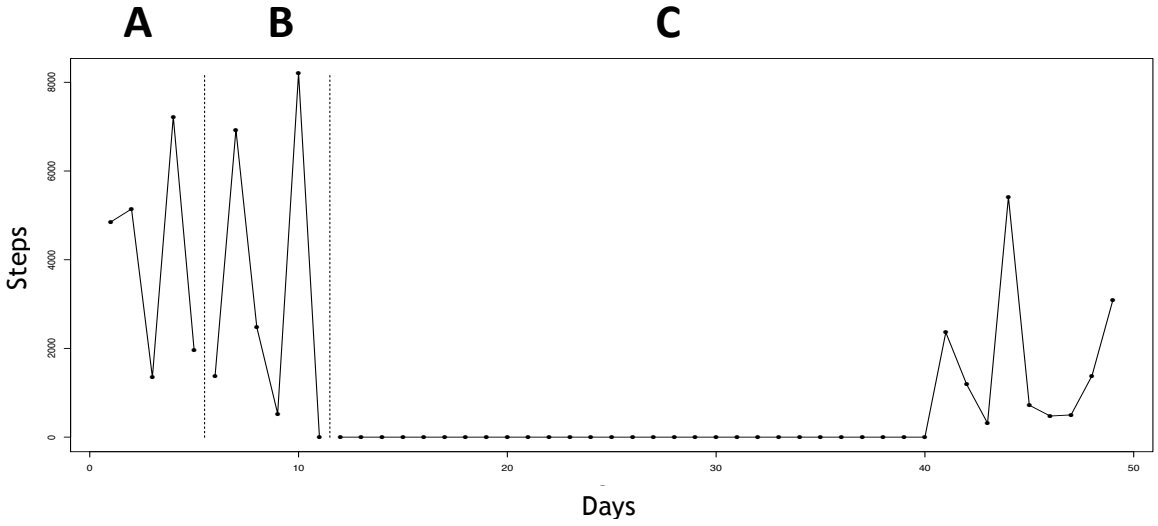


Figure 11: A user's step count data in phases A, B and C (zero values example 2)
Zero values occurred at the introduction of the second intervention phase ("C"), before then returning to non-zero values at 40 days.

9.11.2 Independent variable: Was the intervention delivered and received as intended? (QI 2.4)

Logs indicated that goals were delivered by the app every Monday as planned; however some users experienced unintentionally low goal values due to the presence of zero count data within the baseline phase (from which goals were generated). Figure 12 shows the number of users who had been exposed to or 'received' the different intervention phases and components. The majority (68.7%, 55/80) consented to step data being collected (which provided data for

phase A) and used the app for long enough to receive both the self-monitoring (phase B) and goal setting (phase C) components[28]. Four users appeared to stop using the app before phase C was activated. Some users (22.5%, 18/80) consented to collection of step data, but did not go on to use the app. Three users consented to the study but did not grant access to their step data. The majority of participants (77.5%, 62/80) used the social comparison feature at least once.

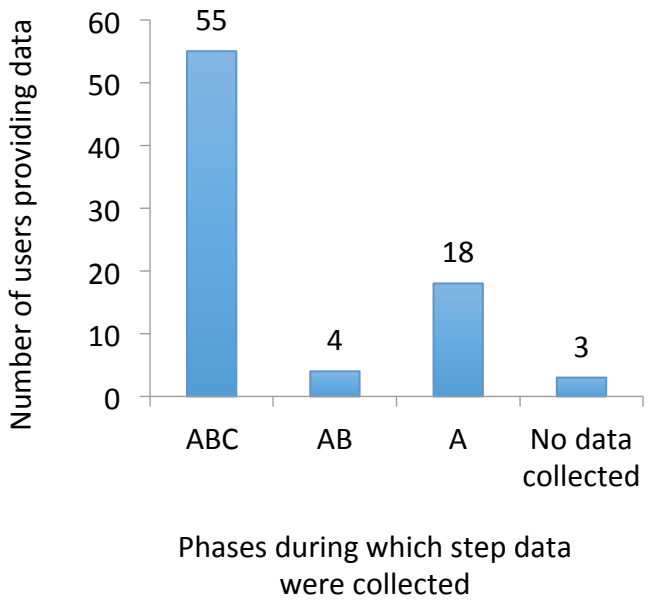


Figure 12: Number of users that received different intervention phases and allowed access to data (including baseline data)

9.11.3 Baseline: Can baseline data be used to predict patterns of future performance? (QI 3.5)

Several users showed high variability in baseline data; examples are shown in Figure 13. Variability appeared to be intrinsic to the dependent variable itself (i.e. step counts appear highly variable), however visual analysis also suggested that intermittent zero count data may contribute to variability. Although users with consecutive zero count values in the baseline phase (e.g. Figure 10) would

[28] While these participants received self-monitoring and goal-setting features, minimum data requirements were met in all three phases for only 43/80 users (53.8%), see preceding section on Dependent Variable.

appear to have highly stable data, these should be excluded from effectiveness analyses.

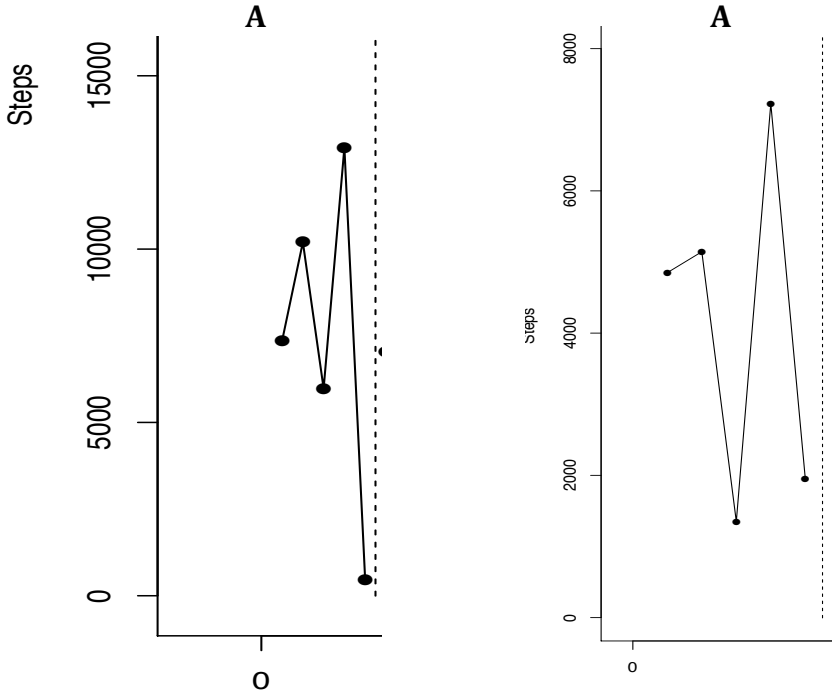


Figure 13: Variability (i.e. low stability) in two example users' baseline phases
The diagram on the left indicates that zero step counts may contribute to data variability, while the diagram on the right indicates that there is also natural variability in step count data

Relative stability (i.e. variability around the mean (Schoenfeld, Cumming et al. 1956)) was calculated for the baseline phases of all users, excluding those whose baseline phase consisted entirely of zeros (n=8). As Table 2 shows, only 13 (18.8%, 13/69) users would be included in effectiveness analysis using a stability criterion of ≤ 5% (which is recommended to demonstrate high levels of experimental control (Schoenfeld, Cumming et al. 1956)). Setting the stability criterion to 10% would still only include less than a third of users (30.4%, 21/69). Most users (88.4%, 61/69) could be included in analysis by relaxing the criterion to 50%, however this would produce high levels of uncertainty. Eight users (13.1%, 8/61) required <50% stability.

Stability criterion	Number of users (accumulative)
5 %	13
10 %	21
15%	32
20%	37
25%	48

30%	52
50%	61

Table 10: Number of users whose baseline data met different stability criterion

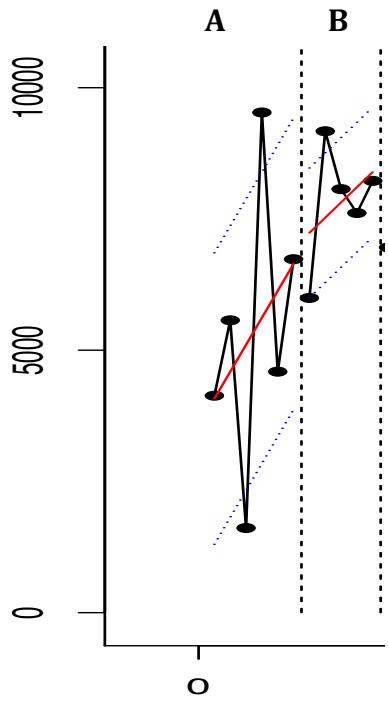


Figure 14: A user's data showing problematic increasing trends in baseline phase A and intervention phase B.

In addition to variability, trends were also found in the baseline phase of some users' data sets. Figure 14 shows an example user's data with an increasing trend line (i.e. where the red 'line of best fit' shows the trend, and blue dotted line shows variability around that trend). Although the trend line predicts future performance, it increases in the same direction as the intended effect of the intervention and thus this dataset should not be used in analysis.

9.11.4 Internal validity: Did the design facilitate at least three replications at three points in time (QI 4.1) and support verification (QI 4.2)?

Figure 15 shows the temporal distribution of downloads (i.e. users' download patterns). The graph indicates that the design facilitated several replications at different time points for phase A to B (i.e. many users downloaded the app at different times) and phase B to C (i.e. users enter phase C on Mondays of different weeks and months).

Multiple baseline verification periods were found to occur naturally within the download patterns of users. Example verification periods are shown within the subsets in Figure 15. The design facilitated verification for a changing criterion design for some, but not all, participants. Log data showed that 47 users (58.8%, 47/80) had the app installed for long enough to implement at least three criteria (i.e. two increases in weekly step goals). Of these users, just over half (53.2%, 25/47) were actually exposed to three criteria; for several others (46.8%, 22/47) the criteria were not changed (as they did not achieve step counts of ≤95% of their step goal).

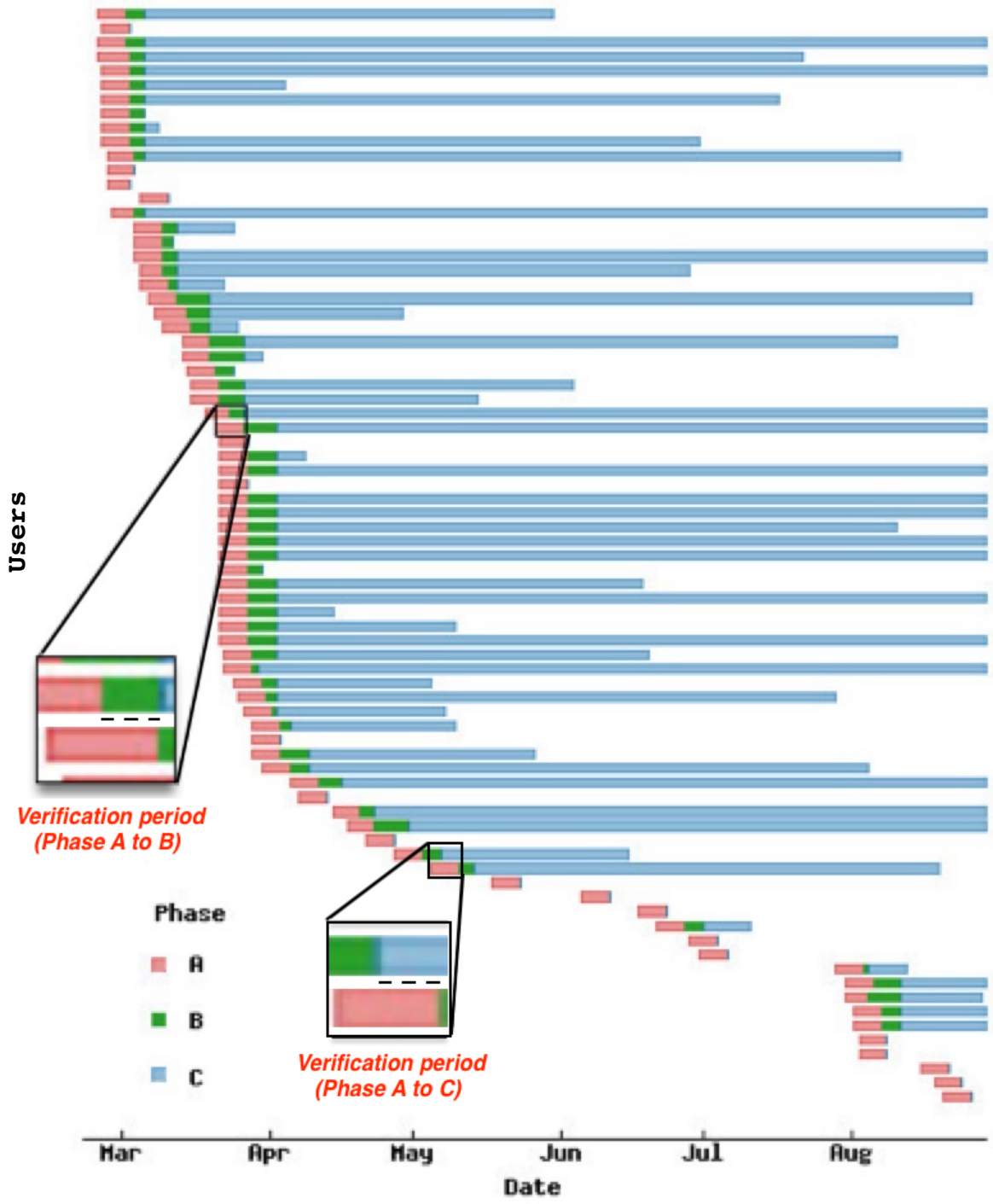


Figure 15: Temporal distribution of user download dates.
This graph shows: the times at which different users downloaded the Quped app from the app store and entered intervention Phase B (i.e. the beginning of the green bar); the time period for which their baseline phase A was retrieved (the red bar), and; the time at which they entered intervention phase C and goal-setting became active (blue bar).

9.11.5 External validity: Was the experiment replicated across different participants and settings, and can we describe them? (QI 5.1, 5.2)

The majority of participants (77.5%, 62/80) entered their gender and age within the social comparison tab. Of those who provided this information, 54.8% (34/62) were male and 45.2% (28/62) were female. Table 11 shows that Quped

was downloaded by participants of different age ranges. Time zone data logs showed that Quped users were both within the UK (62.5%, 50/80) and international (37.5%, 30/80).

Age	All	Males	Females
18-29	24 (30.0%)	13 (16.3%)	11 (13.8%)
30-39	14 (17.5%)	7 (8.8%)	7 (8.8%)
40-49	10 (12.5%)	7 (8.8%)	3 (3.8%)
50-59	12 (15.0%)	6 (7.5%)	6 (7.5%)
60-69	2 (2.5%)	1 (1.3%)	1 (1.3%)

Table 11: Age and gender of participants
(% refers to total number of consenting users, n=80)

9.11.6 Analysis of interviews to explore social validity (QIs 6.1, 6.2)

This section reports the analysis of interviews conducted to explore quality indicators in the SCD Quality Analysis Checklist that relate to the social validity of the Quped study. Themes and associated subthemes are summarised in Table 12. The first social validity quality indicator (QI 6.1) is that the DV (target behaviour) is socially important. From user interviews, one related theme emerged: “the importance of walking”. In relation to the acceptability of the intervention (QI 6.2), a distinction was found between the “acceptability of the physical activity app” (i.e. features to support the intervention) and the “acceptability of study procedures”. Participants’ perceptions relating to these themes and sub-themes are described below. Challenges relating to the operationalisation of quality indicators for social validity are reported for each theme. Methodological implications for exploring social validity when conducting SCD trials are provided in the chapter discussion (section 9.13).

Table 12: Themes associated with social validity quality indicators for the Quped app

Social validity quality indicator	QI	Theme	Subthemes
Is the dependent variable socially important?	6.1	Importance of walking	Improving physical health Improving mental and emotional health Enjoyment of walking Step counts as a measure of walking
Are intervention and study procedures acceptable?	6.2	Acceptability of the physical activity app	Appreciation and interest in tracking steps Perceived effectiveness Technological burden
		Acceptability of study procedures	Informed consent process Data privacy

9.11.7 Is the dependent variable socially important? (QI 6.1)

Four sub-themes were found in relation to the importance of walking: Improving physical health, Improving mental and emotional health, Enjoyment of walking and Step counts as a measure of walking.

9.11.7.1.1 Improving physical health

Some participants felt that walking was an important way to increase their physical health. Participants viewed walking as “good exercise” [P15], a way to “lose weight, get fit” [P7], and the “healthy option” (P12). Because of this, they explained that they were “always concerned” that they “don’t walk enough” (P9), and that they “try to walk as much as possible” (P15).

Walking was important to some participants because they did not do other forms of exercise. Two described how physical health issues prevented them from doing rigorous activity, but walking felt manageable:

> I have knee problems, so I can’t do that much impact sports.... So walking, going for long walks or speed walking is something I can do where I’m feeling actually being active but it’s not hurting me. (P16)

> I had a knee operation a couple of years ago and the doctors say not to run... I wouldn’t go out running for pleasure or fitness. Not running, my knee won’t have it (P12)

One participant described how having a baby had prevented her from doing any other forms of activity:

> "... was the thing that fitted with my lifestyle because I got the baby and that really the only physical activity I could do, walking with him" (P8).

Other participants felt there were "more effective ways" (P14) than walking to improve their health. One felt going to the gym was more important and so did not "care" about walking (P1), while another highlighted that the importance of walking changed depending on the seasons, stating that their "main cardio in the summer time is cycling" (P13).

9.11.7.1.2 Improving mental and emotional health

Some participants noted that walking was important not just for physical health, but also for mental health. One participant explained that health benefits from walking were a "bonus", because they walked primarily to clear their mind as opposed to "solely for physical health" (P14):

> I think walking for me is more of a mental health thing. Rather than a physical health thing... I was going through a bit of a hard time, last semester, and a lot of the ways I would deal or not deal with that depending on how you look at it, is go out for long walks. Because being in the house I felt very stifled, so I would go out for very long walks so that I could feel better, I could clear my head and not sit just stew in it (P.14)

Participants described other ways in which walking influenced their mind-set and mood. One person described how, when walking, they felt they were "doing something constructive and working towards something" (P8).

9.11.7.2 Enjoyment of walking

Participants varied in whether they found the act of walking mentally engaging and fun. One participant noted that they found walking and listening to music simultaneously enjoyable (P15). Another, however, felt more "passionate" about other forms of physical activity and described walking as "boring":

> it's just so boring you know, there's so many things you can do, like ride a bike, and go play tennis and things like that **P1**

Overall, participants felt that it was important to be physically active for physical, mental and emotional health, but varied in whether they thought walking was important to them in achieving health.

9.11.7.3 Step counts as a measure of walking

All participants were clear about the link between step counts and walking. Some monitored their steps to understand if they were meeting a specific 10,000 step goal:

> in terms of steps I think it's just, because I've got a very sedentary work day. It was an idea of whether I am doing that magical 10000 steps a day (P4)

> I believe it's something, they say you do 10000 steps a day, It's the healthy option or something, so I always try and do that (p12)

Others felt other measures beyond steps were important to "capture" (P2) amount of walking, such as distance ("not just how many steps but how far you travelled", P11). Some participants felt that their interest in steps was limited. One participant noted that they "don't care that much" (P8), while another noted they "don't have any particular interest" (P3) in their steps.

Overall participants varied in whether steps counts were important to them.

9.11.7.4 Operationalisation considerations

The finding that not all participants felt the dependent variable was socially important (i.e. not all were motivated and interested in walking) suggests that some of those interviewed would not normally use a physical activity app if it were not for the trial. When using qualitative data collection methods to support quality indicators (i.e. provide evidence) of social validity it is important to acquire diverse perspectives, however researchers should be aware of participants' motivations for taking part in the study and consider how this may influence their perspectives on social validity.

9.11.8 Is the intervention acceptable? (QI 6.2)

In relation to the acceptability of Quped as an app-based intervention designed to improve physical activity, three sub-themes were found: appreciation and interest in tracking steps, perceived effectiveness and technological burden.

9.11.8.1 Appreciation and interest in tracking steps

Some participants felt that monitoring steps motivated them to move more. One noted how steps would "spur" them "to think, you know. You should walk around for a bit." (p4), while another said steps had "always been something I've looked at in terms of... making sure I move around" (P3).

Other participants used Quped to track steps out of interest. One noted they looked at the steps because they had "always wondered" how many they take in a day (P10). Another explained that they did not necessarily use Quped to "do better" in improving physical activity but "enjoyed knowing" their step counts as a "fun fact" (P14).

Beyond daily totals, some participants were interested in seeing how their steps changed over time. One participant noted how they tended to "keep records" of activity in general, including times taken when running (p4) and described being interested in "whether I'm moving less than last year" (P4). Others felt specific days were interesting:

> ... what I found more interesting was actually being able to track back in the history and be able to see when I started the app what was done on what days (P6)

Another participant who was a council refuse collector wanted to find out how his steps varied depending on where he was working each day as "Because some days are slightly heavier than others, depending on the area I'm doing" [P6].

Overall while some users tracked steps with Quped in an attempt to improve their physical activity, others were more driven by curiosity.

9.11.8.2 Perceived effectiveness

Participants tended to describe whether they felt specific app components were motivating and made them more physically active. Some participants felt that goal setting motivated them to walk more as it was an "incentive... I'm going to go out and do that little bit extra" [P12]. One participant described how they particularly liked that goal setting was automated as opposed to setting their own goals:

> I think it's more useful because you're really seeing how you are achieving a certain amount. Like I could set my own goal, but I wouldn't know like.... I wouldn't say this is the goal for next week, like a little bit more... you really don't have anything to compare to (P9)

Other participants felt that Quped's acceptability was limited because the goal-setting feature was not effective. Some described the app as not motivating or "exciting" enough and there was a need for other goal-setting features, such as more interactivity or "points" (P1). Some participants felt that the goals continued to increase beyond what they could manage. One participant felt that they "didn't particularly try to meet goals because they were too high" (P18). Perhaps, most crucially, however, was that some participants felt the increasing goal-setting feature negatively affected their emotions. A participant explained that they could not "keep up" with the ever-increasing goal and found this "depressing":

> after a certain point I felt too much.... it's kind of depressing after a while ... it felt good to achieve something to begin with, but then after a certain point it became, I can't keep up with this. I'm feeling I'm not going to manage (P8)

Another participant described how "It's not nice knowing you're not meeting your goals" that it "doesn't make you feel good" (P10). As such, this participant felt that people should have more control over the goal-setting feature (i.e. the ability to "turn it off" when they were not having "a good day" (P10)). Another participant highlighted how the increasing nature of the goals could be "addictive" to the point where it becomes harmful to physical health:

> ... you could get someone who is obsessed... it must be unhealthy to walk so much... there are people out there... who would download

> this, who are susceptible to that addiction so, I think you really need to be careful... I've got a bad history of doing too much exercise, like I was addicted to exercise at one point, and that's why I think now I'm really careful (P15)

However, the participant went on to say that for this reason they liked Quped because it was not too motivating:

> it's motivating but not to the extent that I was overdoing it P15

Overall, goal setting was reported to lead to unintentional negative effects on peoples' emotions. The acceptability of Quped's social comparison feature was also not perceived to be particularly high. Some participants reported feeling motivated if they saw that they were "doing better than others which makes you feel like you're healthy" (P8). Others recounted being "below" the average for their age range: for one participant this was found to be motivating "to try and catch up" (P11)". However, another felt that although it "gives you an idea that you're not doing enough" they did not know "if it necessarily made me want to get up" (P4).

Some participants did not use the social acceptability component because they found it more useful to compare themselves to their own goals:

> I think I always view my fitness as a personal challenge about me, getting fitter than I was last week (P13)

Other participants felt the social acceptability feature did not provide them with enough data to compare themselves to others. One reported that at one point there was "literally no data" shown for some age groups (as no one from this age group had downloaded it at that time) (P2). Others felt the feature should provide information on other people's physical ability to allow them to interpret their step count. One described how providing only other people's age and gender meant they could unknowingly be comparing themselves to "a community of athletes" (P5), while another felt their own physical health issues meant that social comparison data was unlikely to be "relevant" to them, which was "disheartening" (P10).

9.11.8.3 Technological burden

Participants felt Quped was not intrusive or burdensome in using phone battery or memory. One participant mentioned that they were often concerned about the influence of apps on their phone battery levels but running Quped did not result in any "battery decrease" (P5), and another mentioned that their battery had not "drained at all" (P18). In relation to phone memory, one noted that they did not want to allocate any to Quped, but that this was because they already had very limited space (P14), while another was "super OK" with Quped running on their phone memory, because it was a "study project" (P5).

There were mixed feelings on the weekly notification Quped provided to inform users that a new goal had been set. One participant said they would have liked daily (rather than weekly) goal notifications (P11), however, others were less positive. One participant noted that because they "didn't like" notifications for apps in general ("I always turn them off"), this may have lead her to turn off notifications for Quped from "the beginning" (P1). Similarly, another mentioned she did not often receive notifications from Quped because she often left her home at phone due to feeling notifications from other apps were "overwhelming" (P7).

Finally, one participant described how at time she felt "frustrated" with Quped because it needed to be continuously carried to accurately represent their steps ("you have to have your phone in your pocket for the steps to get counted" (P16)", which she could not do while at work as a waitress.

Overall, the intervention was felt to be acceptable although not necessarily effective, and participants highlighted how the app could negatively affect some users with health issues.

9.11.8.4 Operationalisation considerations

This section highlights the complexity of social validity: users may value an app for reasons beyond those anticipated by researchers (e.g. because they find certain features interesting, rather than because they feel the app improves their physical activity levels). This underscores the importance of using qualitative open-ended approaches (as opposed to closed and pre-specified

survey items) for capturing aspects of social validity that are relevant to participants. Another important finding is that some participants found aspects of the app "disheartening" and detrimental to their wellbeing. This highlights the need for social validity quality indicators to address specifically any adverse effects for users engaging with a physical activity app.

9.11.9 Are the study procedures acceptable? (QI 6.2)

In relation to the acceptability of the study procedures, two sub-themes were found: informed consent process and data privacy.

9.11.9.1 Informed consent process

To support the in-app consent process, a participant information sheet was launched upon installing the app. This contained text informing users that data would be collected for research. Some participants felt that this information was important. Interviewees felt it was "good to know exactly what data is being collected" (P10) and that the information had increased how "comfortable", "confident" (P5) and "safe" (P1) they felt taking part in the study.

On reflecting upon the clarity of the information provided, some participants described the participant information as "quite clear" (P4) or "very clear" (P11) and that it "made sense" and was "understood" (P16). However, participants wanted more information, including an explanation of how their step counts were collected from before the app was installed. One participant described how they were "surprised" by this and felt it could have been better introduced within an in-app "welcome message":

> Immediately there was some historic data there, and that surprised me, and I guess it's drawing on historic data that's on my phone, whereas I was thinking I only just installed, how come there is historic data there? Erm, so a little message to say welcome to Quped, here's data from before you discovered us kind of thing. And we'll track you going forward, might have been a gentle introduction erm, for the situation (P13)

Another participant explained that they had wanted "advice" in the app and "educational" health information, including what is considered "normal" and

healthy (P1). They also wanted the participant information sheet to provide more information about "the actual study" and "why it is you're doing it" (P1).

Participants varied in whether they actually read the information sheet. Some claimed that although they had done so, they did not think others (P11, P10) or the "average person" (P3) would, and likened it to the "small print" (P2, P10, P11). Other participants recounted scrolling through the information to consent to the study without reading it, and how they were "conditioned to check the box" (P3). They admitted that they rarely read these details for any app (P11, P16), suggesting this information generally is not important to them. Another felt that they "should" read this information even though they do not (P17).

Some participants had read the participant information sheet but could not remember the content or process of consenting. Interestingly, they felt this was indicative of it being acceptable. For example, one participant described how the overall process of reading and consenting to the study must have been easy and "seamless" because they could not recall it:

> It must have been quite seamless otherwise I would have remembered... you tend to remember bad experiences with apps rather than good ones... so I think it must have been fine". P18

Thus, the information sheet was perceived to be clear and acceptable to those who had read it, although some participants felt that more information should have been provided about the study and the data collection process.

9.11.9.2 Data privacy

Most participants felt that the type of data collected by the Quped app for the study was acceptable. This included step counts, which participants described as "not secret" (P8) or "not private" (P9) or "just steps and so I'm happy to share" (P8). Highlighting the link between the acceptability of the intervention and acceptability of experimental procedures, one participant noted that he was fine with steps being collected, as this was required in order to receive a step goal (P9). Participants were also fine with age and gender being collected.

In relation to user logs, participants felt records of how they had used and interacted with the app were not "secret" or anything to be "embarrassed" about (P9). Others explained that this was because of the nature of the app, as it was "neutral" (P5), and "very 'PG'" (P14), indicating that because there was no explicit content, they would feel no "shame" (P14) in others knowing their usage patterns. They also felt the fact that the app was developed for research purposes validated data collection: one described how they were more "confident" because "someone else cares about my privacy" (P5).

Quped collected time zone data that identified whether or not participants were likely to be in the UK. Many participants said they would not be comfortable if detailed location data was collected (this would be 'intrusive' (P18) and 'weird' (P15)). For example, one felt concerned at first that the app could track bathroom visits but felt more "comforted" after reading the participant information (P5). Another participant pointed out that some people might automatically assume detailed location data is collected for a walking app (P9). This participant, whose home country was not the UK, also noted that the acceptability of collecting location data might vary by country – while they felt "safe" in the UK, they described how in their home country "it's the kind of information someone could use to rob you" (P9).

9.11.9.3 Operationalisation considerations

Asking participants about the acceptability of specifically *study procedures* (as opposed to the Quped app itself) revealed considerations for designing and operationalizing the consent process, and data collection process. Specifically, researchers should ensure that sufficient study information is available and accessible "in-app" to those who are interested, and ensure that the app provides clear descriptions of how and when data will be collected and used. Furthermore, while users found the data collected in this trial acceptable, it will be important to consider the risks and acceptability in collecting any location data for an SCD trial.

9.12 Effectiveness considerations

The work in this chapter reports on the use of the OSDAS Framework to design a SCD study within the app store deployment of a physical activity app. The research question was not whether the app is effective, but whether a single case study conducted over the app store would be of sufficient quality to conduct a robust effectiveness evaluation.

To progress from the OSDAS Framework towards effectiveness analysis, researchers would use findings from stage three to either i) return to stages 1 and 2 of the OSDAS Framework to redesign the app and logging architecture to optimise the data collection process, or ii) select those who provided suitable data and proceed with effectiveness analysis.

In the Quped study, baseline data was found to be highly variable and revealed that in order to include the majority of users within effectiveness analysis, a stability criterion of 50% variability around the mean would be required. This high level of variability would reduce confidence in the effectiveness results. There were also many zero-value step counts within different phases, which may also require excluding users. Finally, researchers would need to decide whether to include only users whose data patterns facilitated verification (QI 4.2) (e.g. users with overlapping experimental phases), or use these participants to verify effectiveness results for the entire data set. The latter is a more liberal approach that enables more users to be included in effectiveness analyses.

9.13 Discussion

This chapter presented the application of the OSDAS Framework to the design and deployment of one app store physical activity app, Quped. Testing the framework involved assessing the extent to which Quped met SCD criteria. Of 25 criteria quality indicators identified, most (84%, 16/25) were at least partially operationalised; the majority (64%) were fully operationalised, four (16%) were partially operationalised, and four (16%) were not operationalised.

Full operationalisation was inhibited by: the nature of app store (app store policy or the remote nature of the trial); technical feasibility (e.g. requiring

participants to own equipment beyond in-device sensors); and data privacy concerns (users sending identifiable data remotely). Data quality in the Quped study was limited by zero-value step counts and data variability.

9.13.1 Refining the OSDAS Framework

The four quality indicators that could not be operationalised in the design and deployment of the Quped app should be *considered* when designing SCD studies for app store settings, but their operationalisation should not be an essential requirement. This differentiation between "essential" quality indicators and those "to be considered" means that researchers can begin the crucial task of evaluating the effectiveness of app store apps, on the basis of what is currently possible whilst being aware of potential weaknesses in the research design. Having criteria "to be considered" ensures researchers will continuously strive to operationalise these whenever possible (including when technological advances and new app store policies enable new ways of operationalizing these).

The following sections discuss implications for labelling the four indicators that were not operationalised as not "essential", but "to be considered". Specifically, the following sections outline the impact of these quality indicators on internal validity (i.e. confidence in any conclusions relating to effectiveness), and any further research needed.

9.13.1.1 QI 1.5: In the case of remote data capture, the identity of the source of the dependent variable should be authenticated or validated

Dallery et al (2013) recommend authentication within their criteria for mobile-supported SCDs. It is surprising that this criterion could not be easily operationalised given it is designed for remote smartphone studies, but this finding highlights the additional challenges raised by fully remote trials, as opposed to those that incorporate both remote elements and face-to-face contact. Methods and technologies required to authenticate data, such as devices that record heart rate or fingerprints (Cornelius and Kotz 2010, Israel et al. 2005, Wiederhold and Wiederhold, 2005), were not possible to embed within Quped, and could not be provided remotely to participants. Requiring users to own these devices to take part in the study would drastically limit the number of

participants. Furthermore, requiring users to send such personal and identifiable data remotely to researchers, without significant efforts to enhance data security, would raise data privacy concerns. Consequently, it was unknown whether the source of the data (DV) was indeed the smartphone owner, as opposed to a friend borrowing their device, for example. HCI studies have noted that activity tracking can be a social activity (Maitland et al. 2006, Maitland and Siek 2009) and qualitative data has indicated that some individuals share data and devices (Rooksby et al. 2014). If a substantial number of participants gave their device to friends who were more active, this would reduce internal validity and erroneously increase the apparent effectiveness of the app. Therefore, researchers should consider this quality indicator when implementing the OSDAS framework. More research is needed to further understand device sharing behaviours, low-cost ways to objectively measure or detect device sharing, and how to increase data security when transmitting sensitive personal details to and from researchers in an app store setting.

ESSENTIAL			
Criteria Group	QI no.	Quality indicators (QI)	Data collection required?
Dependent variable (DV)	1.1	DV is described with operational precision and is measured with a procedure that generates a quantifiable index	No
	1.2	DV is repeatedly measured over time, at regular intervals (i.e. with equal increments between each measurement).	No
	1.3	Sufficient number of data points are collected within baseline and intervention phases (minimum of three/five)	Yes
	1.4	Data is collected or referenced on the validity and reliability of dependent variable measurement	No
Independent variable (IV)	2.1	IV is described with replicable precision	No
	2.2	IV is systematically manipulated and under control of experimenter, and is continuously implemented over time (Changing criterion)	No
	2.3	If multiple treatments or intervention components are examined, each component is introduced separately	No
	2.4	Fidelity (delivery and receipt of intervention) is measured	Yes
Baseline	3.1	Baseline phase that provides repeated measurement of the DV is included	No
	3.2	Baseline conditions are described with replicable precision	No

	3.3	Baseline lengths vary across participants (Multiple baseline)	No
	3.4	Baseline establishes a stable pattern of responding	Yes
Internal validity	4.1	The design provides at least three replications of experimental effect at three different points in time	Yes
	4.2	The design controls for common threats to internal validity	Yes
	4.3	Participants are randomised to different experimental sequences (e.g. baseline and intervention phase lengths)	No
	4.4	Participants and assessors are blinded to the phase of the intervention	No
External validity	5.1	Design supports replication of the experiment across participants and settings	Yes
	5.2	Participants and critical features of the setting are described with sufficient detail (e.g., age, gender, health condition, therapeutic setting).	Yes
	5.3	The process for selecting participants is described with replicable precision.	No
TO BE CONSIDERED			
Criteria Group	QI no.	Quality Indicators (QI)	Data collection required?
Dependent Variable	1.5	In the case of remote data capture, the identity of the source of the DV is authenticated or validated	Yes
Baseline	3.5	Baselines are independent (Multiple baseline)	No
Internal Validity	4.3	Participants are randomised to different experimental sequences (e.g. baseline and intervention phase lengths)	No
Internal Validity	4.4	Participants and assessors are blinded to the phase of the intervention	No

Table 13: The refined SCD Requirements Checklist (V2)

9.13.1.2 QI 3.4: Baselines are independent

Baselines would not be independent if users installed the app on multiple devices, or installed and reinstalled the app. A previous automated RCT used device IDs to overcome this problem (by linking together data sets from one device across installations) (BinDhim et al. 2014). However, App store regulations have since prohibited device identification in an effort to protect users' privacy, and instead only permit unique IDs to be assigned to an installation. Other app stores, such as Google Play, are also continuously considering ways to safeguard users' data privacy (Rooksby et al. 2016). These regulations will make it increasingly difficult for researchers to ensure baselines are independent, yet it may be possible to design "in-app" features that can assign user IDs, as opposed to relying on mechanisms facilitated by Operating

Systems. This is an important area for further research and design work, as failing to operationalize this quality indicator may risk violating assumptions relating to data independence for some statistical analysis techniques when evaluating effectiveness. As discussed in relation to QI 1.5 above, it will be crucial to consider data privacy and security if researchers are attempting to uniquely identify app store users.

9.13.1.3 QI 4.3: Participants are randomised to different experimental sequences

Randomisation in SCDs is not essential but can improve internal validity (Guyatt et al. 2000, Tate et al. 2013). The relatively few SCD established checklists that required randomisation (Tate et al. 2013, Kratochwill 2014), and others who have commented on the approach (Wolery 2013) noted randomisation for SCDs is challenging in real world settings. Unlike group-based designs, which allocate participants to different versions of an intervention, SCD randomisation involves allocating users to different lengths of baselines and intervention phases. Randomisation could not be incorporated In the Quped study due to how the baseline phase was operationalised (i.e. using 1 week of retrospectively collected data). Yet, other operationalisations such as a 'dormant' baseline (whereby users experience a blank interface or other a non-fully functioning app) may be more amenable to randomisation. Overall there is a great need for HCI researchers to explore the design of baseline phases that users actively experience, yet find acceptable. Future research should also explore ways to incorporate a longer *retrospective* baseline phase than 7 days (i.e. by querying data from the Health App as opposed to directly from the M7 chip), as this would enable greater flexibility in randomising participants to different baseline lengths. Longer baseline phases may also help achieve the added benefit of providing a clearer picture of whether data patterns met other quality indicator requirements relating to the baseline phase (i.e. data stability and trend).

9.13.1.4 QI 4.4: Participants and assessors are blinded to the phase of the intervention

Tate et al (Tate et al. 2013) included blinding of both participants and assessors within their SCD checklist. The authors noted that blinding is not often discussed in behaviour change SCDs and can be difficult to implement (compared to drug

trials which use placebo pills, for example). They state that including blinding items in their checklist "serves as a reminder of their critical role in minimising bias" (p. 625). In the same way, the OSDAS Framework should incorporate this as something 'for consideration'. It was not possible to blind Quped users as they actively experienced the introduction of new app components representing different phases. The difficulties in blinding and using 'placebo' conditions for mHealth technologies has been noted (Torous and Firth 2016, Jake-Schoffman et al. 2017). While blinding app users is currently a challenge that requires further research and design work, it is less difficult to imagine how a trial or even analysis software may be designed to facilitate blinding of assessors during effectiveness analysis (for example by blinding assessors from seeing the experimental phase that a particular set of data was collected from).

Criteria group	**Quality indicator (QI) to test**	**Corresponding SCD Requirements QI number**
ESSENTIAL		
Dependent variable	Are there a sufficient number of data points within baseline and intervention phases	1.3
Independent variable (IV)	Was the intervention delivered and received as intended?	2.4
Baseline	Can baseline data be used to predict patterns of future performance?	3.5
Internal validity	Did the design facilitate at least three replications at three points in time?	4.1
Internal validity	Did the design facilitate at least three replications at three points in time and control common threats to validity?	4.2
External validity	Was the experiment replicated across participants and settings?	5.1
External validity	Can participants and settings be described?	5.2
Social validity	Is the dependent variable socially important?	6.1

Social validity	Are intervention and study procedures acceptable?	6.2
TO BE CONSIDERED		
Dependent variable	Can the identity of the source of the DV be authenticated or validated	1.5

Table 14: The refined SCD Quality Analysis Checklist (V2)

9.13.2 Data problems

Importantly, the data quality issues found relating to zero count values or "missing data", and high variability of data, are two of many that could have occurred (and may in future studies). For example, a greater number of users could have objected to providing demographic information, which would limit external validity, or not have an app installed for the same length of time as Quped, resulting in fewer opportunities to verify effectiveness.

Data issues are likely to depend on the design and user experience of the specific app and logging architecture being assessed, and as such, these data issues are not used to refine the OSDAS checklists. Instead, they provide insight into some of the challenges researchers may encounter in stage 3 of the framework. Upon finding variability, researchers have the option to extend the baseline phase or identify the source of variability to try and reduce it (e.g. by finding whether it is a measurement issue) (Kratochwill et al. 2010). Yet physical activity behaviour can be inherently highly variable across days, or even associated with particular days of the week (Matthews et al. 2002). This can be explored with visual and statistical techniques when analysing effectiveness (Valbuena et al. 2017). In relation to missing data, researchers can use imputation methods to support effectiveness analyses (Smith 2012).

9.13.3 Operationalising social validity quality indicators

Unlike quality indicators relating to the dependent and independent variable, and internal and external validity, social validity quality indicators were operationalised in the Quped study using qualitative interviews (as recommended by the OSDAS Framework). Other qualitative methods may also be

used (such as in-app mechanisms) however, researchers should consider whether and how their choice of method will adequately: (i) explore participants' motivations for taking part in the study (such as open questionnaires, to understand relevant issues of self-selection bias) (ii) understand and address any adverse effects for users in engaging with a physical activity app, or indeed any aspects of the trial (iii) ensure participants understand the purpose of the trial and when and how their data will be collected and used.

9.13.4 Conclusions

The majority of SCD requirements could be at least partially operationalised within the design of a specific physical activity app deployed on the Apple App Store. These findings were used to refine the OSDAS Framework to account for natural constraints placed on SCDs imposed by the nature of real world app stores settings. Specifically, the OSDAS Framework was revised to state whether quality indicators were "essential" or "to be considered". Refining the framework to clarify what is "essential" when designing an SCD trial allows researchers to begin *now* in undertaking the urgent task of assessing the quality of their apps in real world app store settings, to keep up with rapidly evolving technology. However, quality indicators that were not operationalisable in the design and deployment of Quped may be operationalised in other types of physical activity apps, or become more easily operationalised in future (e.g. due to technological advances). Keeping these quality indicators within the OSDAS framework (as opposed to excluding them) means that researchers can continue to strive to improve the quality of automated SCDs in future.

The next chapter draws upon the findings from this chapter, and that of chapters 4, 5 and 6, to outline the implications of the work in this thesis. This includes further research that is needed to improve the implementation of criteria that were "partially" operationalised to further enhance the quality of automated SCD trials facilitated via app stores. Furthermore, the OSDAS Framework described in this chapter was designed for researchers; findings from the thesis as a whole are used to outline future steps that would be required to tailor the OSDAS Framework for use by industry professionals.

Chapter 10 Discussion

The aim of this thesis was to explore the use of rapid research designs and efficient data collection methods within evaluations of physical activity apps and wearables. Three studies were conducted that explored: the extent to which these are being used by researchers (Study 1); researcher and industry professionals' experiences and perceptions in using rapid and other research designs (Study 2); and the development of a framework to support the use of rapid research designs, specifically single case designs (SCDs) (Study 3). The objectives and main findings of each study are summarised below.

10.1 Summary of studies

Study 1 was a scoping review conducted across health and HCI disciplines to describe the extent to which rapid research designs and efficient data collection methods (i.e. in-device sensors and user logs) are used to assess the effectiveness, engagement and acceptability of physical activity apps and wearables (objective 1). Despite the number of studies evaluating these technologies almost doubling within two years, very few employed rapid research designs: one study used an SCD and another used the MOST approach. The majority of studies used RCTs. While most studies used in-device sensors to assess physical activity, those that employed RCTs were more likely to employ research-grade devices. Just under a third of studies assessed effectiveness, engagement, and acceptability together in one study. User logs were frequently used to assess engagement but not acceptability.

To understand how the scoping review findings relate to current practices in academia and industry (objective 2), Study 2 involved conducting 15 semi-structured interviews with health and HCI researchers, industry data scientists, and other industry professionals (CEOs and product designers). Researchers across disciplines and industry data scientists recognized RCTs and research-grade external sensors as the 'gold standard', but debated their suitability for evolving behaviour change technologies. Health researchers valued logging software but required technical support from other disciplines. Health researchers also highlighted the complex relationship between engagement and acceptability, and felt more guidance was needed on what to measure.

Interviewees highlighted the strengths (e.g. external validity) and issues (e.g. ethics) of using app stores to facilitate pragmatic RCTs. Industry professionals described rapid and agile research designs (rapid RCTs, A/B testing and "staggered release") and innovative ways to assess acceptability, which were not detected within the scoping review.

The scoping review findings were found to mostly to reflect experiences of academic researchers meaning that little was known about industry professionals' experiences in evaluating effectiveness (objective 3). Furthermore, it was of interest to explore *why* rapid research designs are not currently widely (objective 4). Therefore, a deeper interview analysis was conducted in Study 2 (using the COM-B model (Michie et al. 2011) to characterize barriers and facilitators to assessing effectiveness of behaviour change apps and wearables and using rapid research designs. An emergent finding was that the industry professionals interviewed were not motivated to assess effectiveness. The perceived potential for effectiveness evaluations to boost company reputation and do 'social good', competed with ultimate priorities to increase profit, and the perceived risks, time and resources involved. Interviewees believed that trial automation could reduce the time effectiveness evaluations took to conduct.

Researchers were motivated to use rapid research designs but felt that these were unacceptable to the scientific community, who would not fund studies that used these designs nor accept them for publication. Furthermore, researchers felt they lacked the time and capability to learn how to implement rapid research designs and statistically analyse results, and described the need for education and supportive tools.

Given that few studies used rapid research designs, researchers appeared to require support to do so, and opportunities pointed towards automated and app store facilitated trials, Study 3 involved developing a framework support researchers in Operationalising Single case Designs for physical activity Apps Distributed via App Stores (the OSDAS Framework) (Objective 5). This involved collating SCD criteria and quality indicators across established guidelines and standards, and testing the extent to which these could be designed within a physical activity app store app and whether data collected would meet

standards. Findings were used to refine and tailor the OSDAS Framework to account for app store characteristics (objective 6): some SCD requirements could not be operationalised due to App Store policies, technical feasibility and data privacy issues. Non-operationalised requirements were added in a separate section the OSDAS Framework; quality indicators "to be considered" by researchers, as opposed to those that are "essential". Overall, the resulting framework has 3 iterative stages (Design, Data collection and Data analysis) supported by a 23-item SCD Requirements Checklist and a 9-item SCD Quality Analysis Checklist.

10.2 Challenges and opportunities in using rapid research designs to assess the effectiveness of physical activity apps and wearables

10.2.1 Rapid research designs

A review of effectiveness evaluations of clinical apps found that rapid research designs were not often used (Pham et al. 2016); the scoping review findings suggest that this phenomenon extends to evaluations of physical activity apps and wearables. This is surprising, given that the types of data these devices produce can support some of the requirements of these research designs (such as the frequent data collection required for single case designs (Horner et al. 2005) and microrandomised trials (Liao et al. 2016). Crucially, findings from Study 2 suggest that the uptake of rapid research designs will not simply improve over time as academics become more aware of their existence and their benefits. The interviews indicate that simply being motivated to use these is not enough: academics potentially lack the opportunity and capability to use them, and solutions to address these barriers are needed.

The interviews suggest one major barrier to using rapid research designs is that RCTs are currently the gold standard in health research. Interviewees across disciplines were aware of this standard, and perceived RCTs to be the only design acceptable to those responsible for approving grant proposals (i.e. funding bodies) and accepting publications (i.e. editors and peer-reviewers). Thus, even if academics evaluating mHealth interventions feel there are more suitable designs than an RCT, they may continue to use RCTs to reduce the risk

of not being funded or published. The stakes would be high: taking such risks could jeopardise their academic career.

Decision makers (i.e. funding bodies, editors, peer-reviewers) should be targeted to ensure rapid research designs are welcomed. Encouragingly, funding initiatives have called for studies that use innovative evaluation approaches to evaluate mHealth technologies (Riley et al. 2013). However, these (and similar initiatives from journal editors) may not be enough to promote sustainable and permanent change in perceptions of rapid research designs, given how deep-rooted the RCT is within academic culture. Evidence is needed that studies implementing rapid research designs produce useful and scientifically rigorous results.

A surprising barrier for academics in using rapid research designs was that learning how to use these research designs, and the experimental set up involved, were perceived to be too time-consuming. The OSDAS Framework was designed to support researchers in conducting research projects efficiently when using rapid research designs in a number of ways. The accompanying OSDAS SCD Requirements Checklist, which draws upon multiple sources to outline SCD criteria, eliminates the need for individual research teams to assimilate these themselves. A set of requirements that can be "re-used" across different projects can improve efficiency when designing software (Toval et al. 2008) such as apps. The operationalisation work outlined in chapter 9 not only ensures that SCDs can actually be implemented for app store apps (and is thus a worthwhile approach), but reports how SCD requirements were operationalised. Hekler and colleagues (Hekler et al. 2016) argue that researchers should "share" how they operationalised behaviour change techniques in an app to promote efficiency, as successful operationalisations can be used in other projects. In the same way, research efficiency can be improved by sharing operationalisations of rapid research designs.

The OSDAS Framework was designed to improve the pace of research in general by supporting evaluations of effectiveness in the real world setting. This allows researchers to understand and account for factors such as engagement that could mediate or influence effectiveness (Glasgow et al. 2003), and thus facilitate a more efficient "translation" of findings to real world settings (as

opposed to highly controlled lab based studies which can produce results of limited relevance to real world settings).

Despite the potential for automated app store trials to improve the efficiency of effectiveness research, some findings from Studies 2 and 3 pose interesting ethical challenges with this approach. Study 2 interviewees felt that it should not be used with more vulnerable populations due to logistical issues they had experienced in acquiring informed consent from patients who required doctors to be present (as part of standard research procedures in the USA). While the OSDAS framework is for use with physical activity apps, which are considered to be relatively low risk, social validity findings in Study 3 revealed that undergoing the trial of the Quped physical activity app could be detrimental for some users. For example, participants described feeling adversely effected by not achieving their target goal and when comparing themselves to others, and noted the potential for users to become obsessive about goals. While this was revealed when exploring social validity in qualitative interviews, none of the checklists informing the development of the OSDAS Framework explicitly addressed the need to report adverse effects. This should be considered as a specific checklist item in future, and more research and design work is needed to operationalise "in-app" adverse reporting to (i) enable users to directly report adverse effects in automated SCD trials, and (ii) provide relevant support for these users.

In addition to the consent and adverse reporting issues that arise when researchers are not in direct face-to-face contact with participants, there are interesting ethical issues in putting yet another app on the app store that has not yet undergone rigorous effectiveness evaluation. One way of protecting users would be to use the "staggered release" approach discussed by interviewees, where the app does not become widely available until found to be effective. Yet, further difficulties will likely arise in using novel experimental procedures that are currently without standard regulation and ethical approval processes (unlike more traditional RCTs). It is not difficult to imagine how researchers and developers may see automated trials as an opportunity to bypass the approval of regulatory bodies such as the Medicines and Healthcare Products Regulatory Agency (MHRA), which can require RCTs and be a slow process relative to app development processes (Vincent et al., 2015). More work is

required directly with the MHRA to shape their policies, processes and guidance in relation to using rapid research designs.

Chapter 6 highlighted the benefits researchers perceived in conducting pragmatic RCTs and using app stores to do so. Beyond greater efficiency, a reported advantage was being able to demonstrate the extent to which users choose to download a particular app. While user selection has been discussed in relation to automated trials of online website interventions (Eysenbach et al 2005), app stores are arguably a more competitive environment, as users are actively presented with an array of apps in the same category (e.g. fitness) competing for their attention. The Quped study attracted only 180 consenting users. While SCD studies require few users compared to group designs (Kratochwill et al. 2010), larger user numbers would support statistical methods that can illuminate for whom the app is and is not effective (Chen et al. 2012, Shadish et al. 2013). Overall, researchers running app store facilitated trials should consider and budget for app advertisement and marketing.

Chapter 7 reported that researchers across disciplines felt more studies should seek to understand how best to disseminate and implement health apps. This can include assessing factors such as adoption, feasibility, fidelity, implementation cost, penetration, and sustainability (Proctor et al. 2011) and reach and uptake (Murray et al. 2016). A small number of studies have assessed reach and uptake for researcher-developed apps distributed via app stores (e.g. apps for mental health; Lattie et al. 2016), however the apps were not designed to support automated effectiveness trials, and thus effectiveness (or whether the data quality could support effectiveness claims) was not assessed. Overall, to take full advantage of running trials in app store settings, further checklist items could be incorporated within the OSDAS Framework to design apps that support "effectiveness-implementation" studies (Curran et al. 2012) that explore outcomes relevant to both in one trial.

Industry data scientists and researchers anticipated challenges with app store trials. One challenge was that the remote nature of the approach (with little or no face-to-face researcher-participant contact) might not be suitable for more vulnerable clinical populations. The ethical implications of app store approaches should be carefully considered. Previous HCI studies have explored informed

consent processes for remote trials (McMillan et al. 2010, Morrison et al. 2012), however health-related apps are likely to require different considerations. Quped user interviews were analysed in Study 3 to explore whether experimental procedures such as informed consent processes were socially valid and acceptable. The results highlight specific concerns that should be addressed in future studies (e.g. to ensure that data collection processes are fully explained in in-app participant information sheets) and the potential negative effects on mental and emotional wellbeing evoked by participating in a physical activity app SCD trial.

Further challenges in running app store trials emerged from putting the OSDAS Framework into practice in Study 3. The removal of non-operationalisable quality indicators means that the framework can be readily used, which is important given the number of evaluations of physical activity apps and wearables which are being conducted but not using rapid research designs. Some non-operationalised quality indicators are unlikely to be operationalisable in the near future (e.g. data privacy concerns are only going to increase). However, it would be useful to assess whether randomisation (QI 4.3) can be fully operationalised and incorporated back into the SCD Requirements checklist: although randomisation is not a requirement for SCDs (Dallery et al. 2013), it could strengthen the internal validity of any effectiveness claims (Kratochwill and Levin 2010). Additionally, to improve the quality of studies using the OSDAS Framework, quality indicators that were 'partially operationalised' should be further explored to understand whether they can be "fully operationalised".

The thesis focussed on SCDs. These are likely to be a useful approach for HCI researchers, as SCDs are coherent with their existing research practices. The scoping review found that many studies employed small sample sizes (almost a fifth with <12 participants), and others have suggested that this is typical of HCI research (Kay et al. 2016). The review also found studies that incorporated similar features to SCDs (i.e. case studies with baseline phases) but had not followed SCD principles. In comparison to other group-based rapid research designs, SCDs can be especially useful for identifying for whom and in what contexts the app works and does not work; this can be used to target users (Hekler et al. 2013), in addition to providing valuable insight for clinical and

behaviour change researchers (Chen et al. 2012). The OSDAS Framework suggests that the SCD approach can be applied relatively late in the design process, once the core components of the app have already been developed (such as features supporting behaviour change techniques). This allows researchers to use the approach to evaluate existing apps, however, it would be most useful to consider SCDs during earlier prototyping stages to allow more flexibility in iterating and improving the app to accommodate SCDs.

Funding bodies and journal editors should not only support innovation, but also designs that are appropriate for specific projects. This requires familiarity with and acceptance of a variety of designs proposed to support efficiency. Different designs are appropriate for different research questions (e.g. MOST might be used to assess the effectiveness of different components), app features (e.g. microrandomised trials might be used for 'JITAI' (Just In Time Adaptive Intervention) apps that adapt to the user over time), as well as stages of the project and settings in which the app will be delivered. Study 3 highlighted how the SCD strengths and weaknesses of SCDs for app store apps. Wyrick et al have shared and reflected upon how they applied the MOST design to a web intervention (Wyrick et al. 2014), for a web-based digital health intervention. They concluded it was a useful approach and considered practical questions that arose when designing their study. More studies are needed which apply rapid research designs to different projects: not only to understand how they can be operationalised, but also their strengths and weaknesses.

10.2.1.1 Are RCTs useful for evaluating app store apps?

As highlighted in the literature review (chapter 2), RCTs are likely to be appropriate when a digital health intervention is 'stable' (Murray et al. 2016) with further changes being unlikely. However, it has been recognised that stability is a relative term without a fixed level that must be reached before an RCT is conducted (Mohr et al. 2015). It has been noted that if changes to the app are not substantial, they may be unlikely to impact the outcome of the trial. Mohr et al (2015) explore the concept of assessing 'intervention principles' in mHealth app RCTs; a change in colour may not influence effectiveness, but a change relating to the behaviour change technique, may do so. Chapter 6 reported that participants experienced difficulties specifically due to continuous

and unpredictable changes in the platforms or 'operating systems' on which apps were based. The effects that operating system updates are likely to have on the intervention content may not be substantial. Overall, this means that using an RCT to evaluate app store apps may not necessarily reduce either internal validity (i.e. relating to which of the different app "versions" are responsible for the change in outcome) or external validity (as the RCT can accommodate the real-world changing nature of apps on app stores). Thus, whether RCTs are suitable for mHealth apps, and those on app stores, depends on how substantial the app modifications are.

There is an important question about whether and when SCDs should be used relative to other research designs, including RCTs. Researchers have debated the relative rigour of SCDs to RCTs. Some researchers suggest that N-of-1 trials are most useful for preliminary testing before an RCT (Kumar et al. 2013). Conversely, withdrawal-based SCDs have been considered the gold standard if the research objective is to understand effectiveness of a treatment for particular individuals (Guyatt et al. 2000, Shamseer et al. 2015). Others suggest SCDs should be used after an RCT (to assess the individual-level effectiveness of an app, and to further tailor the design for specific individuals, once it has been found to be effective "on average", Karkar et al. 2015) or throughout all stages of research from app design through to dissemination and implementation (Dallery and Raiff 2014). Overall, whether researchers solely rely on SCDs to evaluate effectiveness or go on to conduct an RCT may depend at least partially on their confidence in the results produced: this can be conceptualised as a function of whether or not OSDAS quality indicators relating to internal validity (causal claims) and external validity (generalisability of results to other people) were successfully operationalised.

This thesis focussed on the pragmatic challenges (efficiency and external validity) with RCTs In evaluating rapidly evolving apps, however it is important to consider their contribution to knowledge. Traditional RCTs can be limited in understanding "what works for who in what context": they have limited ability to explore the effectiveness of individual components (focussing instead on the entire "intervention package"), and the effectiveness of an intervention for a particular individual (providing, instead, an "averaged" result) (Hekler et al.,

2013). However, the key concept of "realist RCTs" has relevance for mHealth apps. These types of RCT focus on precisely "what works for who in what context", through exploring causal mechanisms. Realist RCTs have been discussed in relation to technologies such as personal electronic health records (Greenalgh et al, 2010) and have focussed on the nature of socio-technological relationships (Coiera, 2004), yet discussions of realist RCTs have yet to make a significant appearance in mHealth literature. More recently, Hatcher & Bonnell have called for greater use of realist RCTs in mHealth research (2016). So far, one study has explored the use of realist SCDs for an app to support long-term conditions (Pham et al., 2017), and more development in this area is needed.

Overall, despite their pragmatic shortcomings; even if an app store app is no longer available, RCTs have the potential to provide definitive evidence of whether an app *has* worked, which can be built upon to inform the development of apps in future. Realist RCTs in particular may actually be "efficient" in relation to their contribution to knowledge from a single trial. Given their focus on understanding what works for who in what context, and the Context-Mechanism-Outcome configurations typically produced using these designs, it would be useful to explore how these trials can be augmented using in-device context-sensing technology to collect and analyse data.

10.2.2 Evaluating effectiveness in industry

Industry professionals felt evaluating effectiveness in general would be time consuming. Although the OSDAS Framework was developed for researchers, it should be adapted to support industry evaluations. This would help to address the problem that many publicly available physical activity apps are not assessed for effectiveness (Bondaronek et al. 2018). However, industry professionals pointed out that they would have to be secure in the knowledge that any in-app automation evaluation procedures would be robust or "bulletproof" in ensuring such procedures did not negatively affect the user experience of the app. This would mean careful operationalisation of SCD requirements within the design phase (stage 1) of the OSDAS Framework.

As well as effectiveness evaluations taking too long, industry professionals perceived them to be costly. They felt, for example, that such evaluations would

require a greater number of data scientists than they could afford to employ. Interviewees also indicated that product designers and CEOs face difficulties (i.e. capability barriers) in setting up and statistically analysing effectiveness experiments. One approach would be to simplify the OSDAS checklists to ensure they are easy to use; Lyon and Koerner (2016) describe techniques for actively simplifying and tailoring implementation processes to suit those delivering interventions (such as healthcare staff). A similar approach could be used so that industry professionals can easily embed experimental procedures in their app.

As an alternative to simplifying criteria, chapters 6 and 7 identified automation and supportive tooling as a facilitator for rapidly conducting effectiveness studies and statistical analyses. Mohr and colleagues' propose that when evaluating mHealth apps in healthcare settings, input from researchers is gradually replaced with automated data collection and analysis (Mohr et al. 2017). In the context of the OSDAS Framework for app store apps, an automated system would reduce the need for data scientists (or academics) to be actively involved in every evaluation. Importantly, the scoping review indicated that within studies that were otherwise comprehensive (in that they measured engagement, acceptability as well as physical activity), use of inferential statistics was limited, thus reducing their potential to advance understandings of overall effectiveness. Therefore, the system could go beyond the OSDAS Framework to support automated effectiveness analysis, and be used by both industry developers and researchers.

While an adapted OSDAS Framework may reduce time and resource costs for industry professionals in evaluating their app store apps, other evaluation barriers identified in Study 2 are unlikely to be solved with a software-based solution. Industry professionals saw little value in evaluating their apps: generating profit was their central concern, and because they did not feel evaluating effectiveness led to increased profits or increased user engagement with their app, it was not deemed worthwhile. Interestingly doing "good" (i.e. improving public health) in evaluating and demonstrating effectiveness was not felt to make a significant impact on profit, yet evaluating a product found to be ineffective (and reported as such in the media) could damage company reputation and reduce profit. However, as noted in chapter 5, industry

professionals may use “staggered” evaluation approaches that only require a portion of users to participate in a trial, thus allowing them to quickly change the product to improve effectiveness and reputation. Thus the relationship between effectiveness evaluations, company reputation and profit may be complex. Overall, evidence of a positive relationship between demonstrating app effectiveness and generating profit is needed. Without this, industry professionals may remain unmotivated to evaluate effectiveness.

10.3 Methodological implications

In addition to research designs, the aim of the thesis was to explore the use of data collection methods that promote research efficiency when evaluating physical activity apps and wearables. This included in-device sensors and device generated user logs to measure physical activity and engagement, respectively.

10.3.1 Physical activity sensors

While there is much to be done to improve the uptake of rapid research designs, the fact that the scoping review found that many studies are already using in-device sensors to measure physical activity is encouraging. In-device sensors improve both efficiency and external validity by reducing the need for face-to-face contact between researcher and participants (Murray et al. 2009), and measuring participants in their own contexts via their own device.

Currently, greater efficiency and external validity in evaluating physical activity apps and wearables requires compromising measurement accuracy. Efforts must be made to reduce the need to sacrifice either rigour or efficiency when conducting evaluations: both are crucial in generating useful research. Studies that validate in-device sensors (e.g. Major and Alford 2016, Duncan et al. 2018) are important, yet these must be efficient in order to support efficient effectiveness studies. This is clearly demonstrated by the OSDAS Framework, which requires that researchers reference (or conduct) validation studies for the instrument measuring the dependent variable. Overall, to improve efficiency, results from validation studies should be readily available to researchers. Greater use of industry-based “research libraries”, such as Fitabase (Mack, 2016), would enable evidence to rapidly accumulate for particular devices, and

researchers could access these to inform decisions when designing efficient evaluations.

In-device sensors are crucial for facilitating automated app stores trials that minimise researcher contact, as they enable continuous measurement of the dependent variable (DV), e.g. step counts as a measure of physical activity. Indeed, Study 3 found that criteria and quality indicators that required researchers to provide users with external research-grade sensors (i.e. for "authentication" via finger prints or heart rate patterns (Dallery et al. 2013) could not be operationalised. Such requirements for SCDs will only be able to be operationalised in app store studies once consumer devices that incorporate these sensors become widely available.

10.3.2 Device-generated user logs

The vast majority of evaluations of physical activity apps and wearables employed device-generated user logs, and interviewees valued their ability to measure engagement objectively. Yet, the interviews revealed that health researchers, in particular, experienced challenges in using logs, and described features that their logging software did not incorporate but would be highly useful, such as individual-level metrics. In addition, setting up logging software to assess mHealth apps required support from HCI researchers and computing scientists. While interdisciplinary approaches are important in advancing mHealth research (Nilsen et al. 2012, Kumar et al. 2013), assembling multidisciplinary teams or requiring health researchers to learn logging skills could slow down a research project. There are therefore opportunities for logging software to become more user-friendly and tailored to health researchers' needs.

In addition to software-based solutions, guidelines and best practices are needed to support health researchers in employing and interpreting user logs. The enormous variety of data that logging software could collect was felt to be comprehensive, but also overwhelming. Outside of mHealth research (in the context of gaming), (El-Nasr et al. 2013) has proposed constraints on logging more data than is needed, such as time and resource costs. Similar pragmatic guidance is needed for behaviour change apps and wearables. A recent

consensus study involving researchers across disciplines (Yardley et al. 2016) also concluded that logs are difficult to interpret, and suggested that qualitative data can be useful to help explain any differences found in log measures between participants (Morrison and Doherty 2014). Beyond this, more research is needed to identify which logs (i.e. user interactions with the device) best 'map onto' engagement to achieve high construct validity. Recommending logs that all researchers should collect will encourage standardisation and enable engagement to be more easily compared across apps.

In addition to assessing intervention usage, complex intervention evaluations should seek to understand individuals' experiences and opinions of an intervention (Moore et al. 2015). While qualitative methods are useful for exploring user experiences in-depth, they can be time-consuming to conduct (especially if one-to-one interviews) and analyse (Yardley et al. 2016). The fact that few studies assessing physical activity apps and wearables used device-generated user logs to assess acceptability (i.e. as a proxy to improve research efficiency) suggests that there are challenges with this approach. Interestingly, in a review of a range of health interventions (i.e. outside mHealth) Sekhon found that most studies used behavioural measures (such as drop out rates) to understand acceptability (Sekhon et al. 2017). Interviewees felt that logs could be a proxy measure but would need to be recognised as such (i.e. inferior to qualitative interviews, which provide a more valid account of acceptability).

Overall, although most studies in the review used in-device sensors and user logs, researchers recounted challenges with these, suggesting the need for tools and guidelines. Crucially, given that fewer than a third of scoping review studies evaluated effectiveness, engagement and acceptability together (scoping review), the ease by which this data could be collected via technological advances should mean that more studies are able to collect data on all three outcomes in one study. Assessing effectiveness, engagement and acceptability together will help to develop understanding of what apps or app components work for whom, and in what circumstances.

10.4 Strengths and limitations

10.4.1 Strengths

Multiple disciplines contribute to mobile health research (Blandford et al. 2018). Therefore, to understand the current state of the field in evaluating the impact of physical activity apps and wearables, this thesis employed an interdisciplinary approach. A range of health and computing science databases were included in the scoping search strategy to provide a comprehensive review of research across disciplines. Interviews were conducted with individuals from multiple disciplines and sectors, which provided different viewpoints on the review findings beyond the authors' own initial interpretations and disciplinary perspective. Furthermore, interviews improved the usefulness of review findings by understanding their relevance to the everyday practice of different disciplines, and identified additional research designs that had not emerged in the scoping review (Arksey and O'Malley 2005). Industry professionals, in particular, described agile and efficient research designs and methods that had not been detected in Study 1.

Established methodological and analytical approaches were employed in this thesis, including the scoping review framework (Arksey and O'Malley 2005, Levac et al. 2010) and qualitative framework analysis (Spencer and Ritchie 2002). The latter enabled in-depth understanding of participants' contexts that could help explain different experiences and perspectives. Use of the COM-B model (Michie et al. 2011) allowed capability, opportunity and motivational factors to be compared across two activities (using rapid research designs and evaluating effectiveness) to understand whether one simple solution could improve the prevalence of both activities.

In addition to understanding the state of the field in using efficient research designs and methods, and understanding possible reasons for the state of progress, the thesis investigated a research approach originating in HCI (the app store approach) that has a large potential to increase research efficiency. This was achieved using established HCI design methods (i.e. gathering requirements and prototyping) to develop and test a framework. The Quped study identified

first hand the challenges researchers would experience when employing the framework, ensuring it is grounded in experience.

10.4.2 Limitations

The scoping review, by nature, did not include any assessment of the methodological quality of the included studies (Levac et al. 2010). The focus on physical activity, engagement and acceptability meant that other important aspects of evaluation, such as reach and uptake, secondary clinical and psychological outcomes, cost-effectiveness, and the statistical analysis methods that studies used, were not reported.

Only 15 participants across health research, HCI research and industry were interviewed in Study 2. Although the contexts of these participants were explored in depth, the full range and diversity in perspectives and experiences is unlikely to have been reached within each discipline or sector.

The focus of this thesis was on evaluation of physical activity apps and wearables, which limits the generalisability of the findings to evaluations of other behaviour change technologies. The OSDAS Framework was explored and refined using an iPhone, which may limit the extent to which quality indicators are operationalisable on other platforms Finally, the checklists used in Study 3 were not identified or collated systematically; some checklists and quality indicators may have been missed. To date, only one of the checklists used (Tate et al. 2013) has been validated as a tool to analyse study quality.

The literature review (Chapter 2) concluded that rapid research designs could address a central shortcoming of RCTs: efficiency. Efficiency was discussed in relation to ensuring research is “rapid, responsive and relevant”. Study 3 explored whether it was feasible to implement an app store-based SCD (including the feasibility of operationalizing SCD experimental requirements and data collection processes, and data quality analysis), it did not explore whether SCDs were indeed “rapid” to implement. The overall amount of time such a trial may take researchers to conduct was not investigated, nor was the remaining crucial research task of analysing effectiveness for SCDs. Effectiveness analysis alone may take a considerable amount of time, especially given researchers’

seemingly limited confidence in statistical analysis methods for rapid research designs (as identified in Chapter 7). Importantly, study 2 revealed that designs such as MOST were felt to be time-consuming to set up, which echoes anecdotal accounts reported by researchers elsewhere (Whittaker et al., 2011). More empirical assessment of efficiency would allow assessment of whether rapid research designs are indeed rapid, relative to RCTs.

In addition to efficiency, the literature review concluded that another shortcoming of RCTs, external validity, could be addressed by using app store trial approaches. While external validity quality indicators were successfully operationalised in the Quped app, the OSDAS framework does not specify precisely which contextual details and user characteristics should be collected. In Study 3, only users' gender and age were collected, which does not take full advantage of the ability of SCDs to assess effectiveness for particular types of individuals' in their own contexts. Interviewees in Study 2 discussed relevant challenges (i.e. asking participants for more details leads to an increased likelihood to drop out) and Quped users in Study 3 noted that they would not have been comfortable had Quped collected location data. More research is needed on how to collect data securely for research purposes and maximise the acceptability of that data being collected.

10.5 Recommendations and future research

The exploratory work in this thesis identified three main areas for future research to increase the uptake of rapid research designs and effectiveness evaluations in industry settings. These included the need for evidence, and new guidelines and software.

10.5.1 Evidence of the comparable rigour and greater efficiency of rapid research designs, and of profitable effectiveness evaluations in industry

From the interviews it is unclear whether the academics interviewed actually submitted funding proposals or publications for rapid research design studies and received rejections. Future research should quantify rates of funding approval and journal acceptance across disciplines (e.g. (Sugimoto et al. 2013) to explore the extent to which acceptance/rejection occurs. However, the fact that

academics' expectations of rejection are a barrier means funding bodies and academic journals should consider promoting the use of high-quality studies using rapid research designs.

Although interviewees were motivated to use rapid research designs, and felt that funders would reject these, interviews in Study 2 had limited ability to explain *why* funders may not accept rapid research designs (i.e. any barriers that funders' face relating to opportunity, motivation and capability). Qualitative interviews with specifically those on funding body review boards are needed to understand and provide evidence of any barriers that they face in accepting these designs. Board members may be senior academics such as those in Study 2, but future studies should specifically focus on their role as funders (as opposed to researchers).

If the barriers that funders experience in accepting rapid research designs relate to limited *opportunities* to do so, strategies to improve funding will need to target these opportunities. If, however, barriers relate to funders' own *motivation* to accept rapid research designs, one approach to persuade them may be to increase understandings of their benefits, relative to RCTs. Such an approach will require education and training on the range of rapid designs and their strengths and weaknesses relative to RCTs. This education and training will also help address any *capability* barriers, allowing funders to feel more confident in judging the risks and benefits of funding research that uses these rapid research designs.

Beyond learning about the potential of rapid research designs, funders may need *evidence* that these can produce effectiveness results comparable to the current gold standard RCT. To challenge the established hierarchy of evidence, researchers have compared results from RCTs to observational studies (Concato et al. 2000) and the same is needed for rapid research design studies. This could be achieved either once enough rapid research designs studies have accumulated, or through studies that actively compare effectiveness results from rapid research designs and RCTs for the same behaviour change app. Importantly, some rapid designs are intended to be conducted in advance of an RCT (e.g. MOST), and so should be judged on their ability to inform, rather than fully replace, an RCT.

In addition to evidence that rapid research designs produce comparable effectiveness results to RCTs, evidence is needed that results are achieved with greater efficiency. Thus, explicit measures of efficiency for comparing mHealth evaluation studies are required. These should include contributory factors to the length of time mHealth evaluations take, such as duration of recruitment, or number of measurements (Pham et al. 2016), as well as some indication of their contribution to knowledge relative to RCTs. Consensus studies are needed to produce standardised metrics and measures of efficiency that are important to funders and academic researchers. Researchers using rapid research designs should then be encouraged to report these metrics to facilitate comparisons of efficiency across studies. This will also help to explore whether rapid research designs are indeed "rapid".

In addition to empirically assessing the relative efficiency of specific experimental designs (e.g. SCDs, CEEBIT, microrandomised trials) to RCTs, studies should gather evidence on the efficiency of using app stores to facilitate evaluations in real world settings. These should also compare challenges in automating different types of research design (such as SCDs versus RCTS), and consider efficiency in relation to the (potentially large) marketing budgets required to compete with other apps. Evidence on the methodological strengths and weaknesses of rapid research designs found to have been used in industry and compatible with app store approaches (i.e. A/B testing, staggered agile designs) should also be explored for their use in academia.

Evidence is needed to persuade industry professionals and those funding their app development (e.g. investment agencies) that evaluating effectiveness is worthwhile. Financial and marketing analyses (e.g. which conceptualise effectiveness evaluations as a corporate social responsibility activity (Sen and Bhattacharya 2001) should assess whether, how, when and in what context evaluations can increase profit. Qualitative research and studies using correlational app store mining techniques (Harman et al. 2012) should assess whether users are more likely to pay for apps that have undergone (or are undergoing) effectiveness evaluations. In addition, future research should explore whether and how industry professionals could be motivated to assess the recently devised concept of "effective engagement". This involves assessing

whether users engage with behaviour change techniques in addition to engaging with a device itself (Yardley et al. 2016), and has been proposed as an important metric for future research.

10.5.2 Guideline development

The thesis findings should inform methodological guidelines that support researchers in using in-device sensors and user-logs to collect effectiveness, engagement and acceptability data for physical activity apps and wearables. Those developing guidelines and recommendations should acknowledge the rapid rate at which new smartphones and wearables become available, and their in-built sensors become increasingly sophisticated (Dobkin and Martinez 2018). Thus, although validation studies for consumer devices are needed and especially in real world free-living conditions (e.g. (Major and Alford 2016, Brooke, An et al. 2017, Reid, Insogna et al. 2017), the static nature of guidelines themselves means that recommendations to use particular devices or measures may become quickly out-dated. Therefore, guidelines should be "device-agnostic". Instead consensus studies should generate an agreed set of principles that researchers should use to select measurement devices. These will likely include accuracy, user burden, impact on engagement and data availability, in addition to other important factors such as user safety and data security (Piwek et al. 2016), and whether devices are widely owned.

Consensus studies have already identified the strengths and weaknesses of a range of engagement data collection methods (Yardley et al. 2016). However further research is needed to produce detailed advice and best practice on how to implement user logs, and which logs to collect for physical activity apps and wearables. Consensus groups should be formed of health researchers, HCI researchers and industry data scientists, and also users who can provide insight into social validity issues in collecting particular data.

More theoretical work is needed to distinguish and define the relationship between engagement (device usage) and acceptability before recommending the use of logs to assess acceptability. Interviewees felt the suitability of this approach depended on the number of users, study aims and app stability (i.e. whether improving and redesigning an app or simply monitoring acceptability for

a highly-developed app). Some aspects of acceptability (e.g. trust) were considered not quantifiable via logs. Thus, guidelines should identify which dimensions of acceptability, if any, can be assessed using logs. Researchers should also explore other methods for assessing other dimensions of acceptability remotely. This might include user-entered text logs (as identified in the scoping review) and methods outlined by industry professionals in Study 2, such as video capturing and app store reviews.

In addition to guidelines for using rapid research designs, the difficulties in performing the scoping review across disciplines (chapter 4) highlighted the need for reporting guidelines for app and wearable studies. Current CONSORT-EHEALTH guidelines for digital health technologies focus on RCTs (Eysenbach et al. 2011) and therefore these should be adapted for different research designs and adhered to as far as possible by different disciplines.

10.5.3 Design and development of software tools

Logging software should be developed to support academics, particularly health researchers, in evaluating engagement with apps and wearables. To identify features that support log analysis and engagement evaluation a review should be conducted of existing relevant software (e.g. app analytics, Google Analytics and those developed to enable health researchers design app interventions (Masters 2014, Aranki et al. 2016). Studies with researchers should be conducted to understand which features and functions the software should support. However, findings from study 2 and 3 can inform some of the requirements for such software. Researchers described the need for a system that can help to explore relationships between engagement, effectiveness, and acceptability, as currently available software only provides aggregate-level data. Beyond collecting data, researchers would like support in cleaning, processing and interpreting data. Therefore, a system should:

- Collect and store: user-interaction logs; sensor data; qualitative acceptability data (qualitative user-entered text logs, or interview transcripts data alongside logs); and various other file-types such as video-interaction logs to support triangulation and understandings of not only “what” users’ behaviours were, but also “why” such behaviours were performed.

- Store data at the level of the individual to enable researchers to analyse relationships between different patterns of engagement, effectiveness, and acceptability
- Provide step-by-step guides to cleaning and pre-processing the data for analysis
- Be easy to use and facilitate exploration of data in relation to particular time periods (e.g. enable comparison of engagement patterns between intervention phases and follow-up phases).
- Provide structured output that can be interpreted by those from non-technical backgrounds

Studies should then be conducted to develop prototypes of the above logging software and assess its usability (e.g. ease in interpreting the data). Future research should also inform the development of software to assist researchers and/or industry professionals in running automated SCD trials. Along with managing log data as described above, this SCD software should:

- Support experimental set-up and data quality analysis.
- Detect data quality issues (e.g. trends and instability or missing data) in real-time (e.g. via a dashboard) to enable researchers to remedy data issues as they occur (e.g. by extending the baseline phase) until all data requirements have all been met.
- Document any modifications made to the app
- Support statistical analysis of effectiveness

Research would be needed to explore the feasibility of the new software (i.e. acceptability, demand, implementation, practicality, adaptability, and integration (Bowen et al. 2009). It will also be important to use ethnographic approaches to understand the everyday activities of industry professionals and thus design any new software to minimise disruption to these practices.

In relation to the OSDAS Framework on which the above system is based, researchers should focus on finding technological solutions to fully operationalise QIs “in-app”, that were partially operationalised in Study 3. In particular, future studies should explore different operationalisations of the baseline phase. Study 3 used historic data (from before the app download) to prevent app users from having to endure a period where the app simply measured activity, with few interactive features. This was not expected to be engaging enough for app store users to continue to use the app. Therefore, HCI researchers should examine ways of making such a baseline phase more engaging. Some studies have used baseline phases with some, but limited, features (Kurti and Dallery 2013, Rabbi

et al. 2015). Baseline phases that are actively experienced by users would provide researchers with greater control to fully operationalise some quality indicators: manipulate the onset of the intervention phase (QI 2.2); vary baseline length (QI 3.3); stagger the introduction of the intervention over time (QI 4.2); and/or potentially enable randomisation to different baseline lengths (QI 4.3).

Other quality indicators that could not be fully operationalised related to describing participants (QI 5.3). Researchers should explore ways to collect demographic data from consenting users that is socially valid, not burdensome to provide, and enhances privacy and protection of personal details and sensitive data. Collecting more detailed data will enable researchers to better understand more specifically for whom apps are effective. Overall, research that improves the operationalisation of SCD requirements will improve understandings of the kinds of app-based interventions that can be assessed using the OSDAS Framework, including different BCTs (beyond goal setting), other health behaviours and Android apps. Designers could also explore whether and how withdrawal designs can be operationalised in app store apps, given their methodological rigour. This work could inform the extension of the OSDAS Framework to support the process of selecting an appropriate SCD based on an app's features and components.

10.6 Contribution to knowledge

This thesis found that rapid research designs are not being used in evaluations of physical activity apps, and the findings also help to explain *why* these are not used (namely, due to opportunity barriers researchers face from funders and other decision-makers, and potential capability barriers in being able to set up and analyse results from rapid research designs). Previous studies have identified that app store apps are not often evaluated for their effectiveness. Study 2 revealed that industry professionals face motivational barriers in relation to assessing effectiveness, and are more interested in engagement. Industry professionals are key stakeholders in the assessment of app store apps, yet their perceptions and experiences were previously unknown. Furthermore, researchers felt more research was needed beyond assessing effectiveness, towards research investigating the implementation and dissemination of apps

(including via app stores). This thesis also contributed a methodological framework for evaluating app store apps using a specific type of rapid research design (SCDs) and found that some experimental requirements were challenging to operationalise in a physical activity app. The thesis contributes ideas for how future research and design work can help towards improving the evaluation of physical activity apps.

10.7 Conclusion

The number of studies evaluating the effectiveness of physical activity apps and wearables is rapidly increasing, yet rapid research designs are not being used. Researchers' awareness of these designs and motivation to use them will not necessarily improve their uptake. Use of rapid research designs may be limited by attitudes deeply embedded in the scientific community that result in decision-makers (i.e. funding bodies, journal editors, peer-reviewers) focussing on gold standard RCT designs and rejecting others. To persuade decision-makers that rapid research designs are useful, more research is needed that evidences their ability to produce effectiveness results comparable with RCTs with improved research efficiency. Complex social opportunity barriers may also be at play in industry settings. Echoing their funding bodies (e.g. investment agencies), industry professionals may not see the value in assessing effectiveness if they do not associate it with increased company revenue. More research is needed to demonstrate the extent to which evaluating effectiveness can be financially worthwhile.

Despite providing rapid results, opportunities to use rapid research designs may be limited by the time required both to learn how to use them and for their experimental set up. The OSDAS Framework was designed to increase efficiency and external validity by outlining requirements for single-case designs (a type of rapid research design) using app stores to automate experimental procedures. However, further work is needed to assess the efficiency of the OSDAS Framework empirically, and to explore its use (and tailor it as required) by industry professionals and for apps targeting other areas of health behaviour change.

The fact that the majority of studies already employ in-device sensors and device-generated user logs to measure physical activity and engagement, respectively, is encouraging. However, methodological guidelines that outline principles for selecting sensor devices and best practices on what logs to collect and how to interpret these are required to support researchers. Reporting guidelines for use across different mHealth disciplines are also needed. Together, these guidelines should promote standardisation, enabling researchers to benefit from the vast amount of relevant research undertaken in multiple disciplines to understand the effectiveness, engagement and acceptability of physical activity (and potentially other health behaviour change) apps and wearables.

Appendices

Appendix 1 Thematic framework for interviews in Study 2

Below is the final set of broad codes and sub-codes constituting the thematic framework that was applied to all transcripts during framework analysis.

Personal/Organisational context	**Assessment**
Area of behaviour change	Sensors
Sector/Discipline	Survey/questionnaire
Position/title	Log data
Geographic location	Interviews/focus groups/qualitative
Project details	Validity and reliability issues
Activities (design/development processes)	Data management and interpretation
Intervention details	Multiple methods/combining methods
	Other methods used
	Reasons for using/not using different measures
Research Design	Justification for assessing/not assessing impact
Experience using RCTs	
Perceptions/knowledge of RCTs	
Experiences using rapid designs	**Future/what's needed**
Perceptions/knowledge/awareness of rapid research designs	Education
Evaluation setting: lab versus field trial	Funding
	Working together/collaboration
Sample size (experiences and perceptions)	Toolkits
	Automation
Information sources about research designs	Focus on implementation
Who decides on research design used	Shared understanding of acceptability and engagement
Participants' explicit reasons for not using rapid research designs	

Appendix 2 Intervention characteristics assessed within studies included in scoping review

Study	Technologies	Key intervention features[a,b]
Walters et al 2010 [40]	Smartphone app + web + SMS	Motivational messages, goal-setting with HCP, counselling with HCP
Kharrazi et al 2011 [41]	Wireless pedometer + web	Interactive personal health record, social network, automated goal-setting, reward badges
Pellegrini et al 2012 [42]	Smartphone app	Group sessions, coaching calls, goal-setting for weight, diet and PA, monetary rewards
Jimenez Garcia et al 2013 [43]	Smartphone app	Visualised feedback and goal-setting, vibration prompts, mood diary
Geraedts et al 2014 [44]	Tablet	Exercise videos and telephone coaching
Recio-Rodriguez et al 2014 [45]	Smartphone app	Automated advice, additional manual input of PA
Clayton et al 2015 [46]	Wearable + web	Share activity profile with HCP, goal-setting with HCP
Cooper et al 2015 [47]	Wearable + wireless scales + web	Online detailed FB, goal-setting for weight loss, online virtual coaching
Granado-Font et al 2015 [48]	Smartphone app	Diet monitoring, Social network
Hurley et al 2015 [49]	Wearable + SMS	Adaptive or static goal-setting, immediate or delayed monetary incentives
Pellegrini et al 2015 [50]	Smartphone app	Diary, goal-setting for diet, exercise and sedentary time
Agboola et al 2016 [51]	Smartphone app + wearable	Tailored messaging, community feature for social support and comparison, goal-setting, educational library, portal for providers
Amorim et al 2016 [52]	Smartphone app + wearable	Face-to-face coaching with HCP, goal-setting with HCP, personalised messages
Duncan et al 2016 [53]	Smartphone app + wearable	Educational materials (guidelines, strategies to promote change in behaviour), goal setting with detailed visual feedback
Jones et al 2016 [54]	Smartphone app + wearable	Motivational messages and self-tracking of daily pain and mood
Ortiz et al 2016 [55]	Wearable	Feedback on PA
Shin et al 2016 [56]	Smartphone app + wearable	Feedback on PA, financial incentives
Taylor et al 2016 [57]	Smartphone app	Feedback on PA
van Nassau et al 2016 [58]	Smartphone app + wearable + web	Feedback on PA and sedentary time, social networking, social game, face-to-face classroom discussions and graded group-based PA led by community coaches
Brickwood et al 2017 [59]	Smartphone app + wearable	Health tips and remote feedback from HCPs, telephone counselling
Ridgers et al 2017 [60]	Smartphone app + wearable + web	Interactive weekly individual and/or team missions or challenges, infographics, motivational videos and

		social forums
Wolk et al 2017 [61]	Wearable	Feedback on PA
Slootmaker et al 2005 [62] [63]	Wearable + web	Online exercise planning and goal-setting
Fujiki et al 2007 .[64]	PDA	Scheduling, automated advice, forum, reminders
Hurling et al 2007 [65]	PDA	Avatar, competitive games, game status notifications
Polzien et al 2007 [66]	Wearable + web	Feedback on PA, calories and weight, online self-monitoring, Goal-setting with HCP
Consolvo et al 2008 [67]	Smartphone app	Mobile wallpaper supports continuous graphical feedback on PA, goal-setting, diary, manually edit detected PA
Faridi et al 2008 [68]	Wireless pedometer	Reminders, share data with HCP
Fujiki et al 2008 [69]	Pda	Avatar, competitive games, game status notifications
Lacroix et al 2008 [70, 71]	Wearable + web	LED lighting provides feedback, online detailed feedback, goal-setting
Albaina et al 2009 [38]	Wearable + interactive photo frame	Graphical feedback, automated goal-setting, virtual coach
Bickmore et al 2009 [72]	PDA + web	Virtual, context-aware coach
Fialho et al 2009 [73]	Wearable + web	Social networks, messaging, accept challenges, user can comment on PA, virtual coach
Arsand et al 2010 [74]	Smartphone app	Automated goal-setting, health information access
Mattila et al 2010 [75, 76]	Smartphone app + wireless scales	Diary, graphical feedback
Penados et al. 2010 [77]	Interactive pocket toy	Mood of toy provides feedback
Lim et al 2011 [78]	Wearable	LED light intensity provides feedback
Shuger et al 2011 [79]	Wearable + web	Feedback on PA, telephone counselling with HCP
Burns et al 2012 [80]	Wearable + web	Wearable LED flashes/changes colour if others physically active
Gomes et al 2012 [39]	Wearable + web	Online game, monetary rewards, friend list, newsfeed
Pellegrini et al 2012 [81]	Wearable + web	Feedback on PA, calories, and weight, online detailed feedback, and goal-setting with HCP
Reijonsaari et al 2012 [82]	Wearable + web	Online detailed feedback, and telephone/online counselling with HCP
Van Hoye et al 2012 [83]	Wearable	Feedback on calories and PA, goal-setting, and HCP coaching sessions
Xu, Poole, et al 2012 [84, 85]	Wireless pedometer + web	Online team game, avatars, message board, teacher customisation
Barwais et al 2013 [86]	Wearable	LED lighting provides feedback, online detailed feedback, goal-setting and motivational messages
Bentley et al 2013 [87]	Smartphone app + Web	Widget shows statistical relationships between diet, PA, and contexts via context-sensing
Chatterjee et al 2013 [88, 89]	Smartphone app + SMS	Motivational messages, personalised newsletter

Fitzsimmons et al 2013 [90]	Wearable	Supports consultations and goal-setting with HCP
Harries et al 2013 [91, 92]	Smartphone app	Avatar, social comparison messages
Hirano et al 2013 [93]	Smartphone app	Wallpaper supports continuous graphical FB, vibration prompts, goal-setting and scheduling
Khalil & Abdallah 2013 [94]	Smartphone app	Social networks, share PA data with team, view team progress
Khan & Lee 2013 [95]	Smartphone app + SMS	Automated advice based on activity levels and location detected
King et al. 2013 [96, 97]	Smartphone app	Wallpaper for continious PA feedback, goal-setting, automated "just-in-time" advice, message boards, avatar
Nakajima et al 2013 [98]	Wireless pedometer + interactive photo frame	Interactive painting provides feedback on PA
Tabak et al 2013 [99, 100]	Smartphone app + web + SMS	Virtual coach, automated advice, social graphs, diary
Valentin & Howard 2013 [101]	Smartphone app	Vibration prompts, avatar, goal-setting
Bond et al 2014 [102] [103]	Smartphone app	Just-in-time audio reminder for PA break, graphical feedback
Caulfield et al 2014 [104]	Wearable + web	Online detailed feedback
Chen & Pu 2014 [105]	Smartphone app	Dyads/friends share badges, different social settings
Glynn et al 2014 [106-108]	Smartphone app	Graphical feedback, goal-setting
Miller et al 2014 [109]	Wireless pedometer + web	Online team game, message board
Thompson et al 2014 [110]	Wearable + web	Online detailed feedback, online PA education, GS, counselling with HCP
Thorndike et al 2014 [111]	Wearable + web	Online detailed FB, monetary rewards, team-based competition
Verwey et al 2014 [112]	Smartphone app + Web	User can enter comments about PA, goal-setting with HCP
Walsh et al 2014 [113]	Wearable + web	User profile, social network, currency based game
Zuckerman et al 2014 [114]	Smartphone app	Wallpaper supports continuous graphical feedback, avatar, automated goal-setting, notifications
Cadmus-Bertram et al 2015 [115, 116]	Wearable + web	Self-regulation skills (goal setting and frequent behavioural feedback)
Direito et al 2015 [117]	Smartphone app	Education, social networking/forums/messaging
Finkelstein et al 2015 [118, 119]	Wearable + web	Monetary rewards, charity donation
Frederix et al 2015 [120]	Wearable + SMS	Goal setting, weekly advice, and exercise training with HCP
Frederix et al 2015 [121]	Smartphone app + wearable + SMS	Tailored exercise training protocols, encouraging messages that change over time texts regarding diet and smoking
Garde et al 2015 [122]	Smartphone app	Team game, reward badges
Gouveia et al 2015 [123]	Smartphone app	Goal setting, contextualizing physical activity via context-sensing, and

		textual feedback that continuously updates
Guthrie et al 2015 [124]	Wearable + web	Goal-setting, rewards (badges, monetary incentives), avatar, social comparison, social facilitation
Komninos et al 2015 [125]	Smartphone app	Audio feedback via degraded music quality until user reaches target cadence
Lee et al 2015 [126]	Wearable + web	Personalised planning and goal-setting
Lee et al 2015 [127]	Wearable	Material engraver; tailored feedback using representative patterns
Martin et al 2015 [128]	Smartphone app + wearable + SMS	Automated coaching texts with positive re-enforcement messages
Munson et al 2015 [129]	Wearable + web	Custom website which links to wearable and social networks, automated goal-setting, public goal commitments
Rabbi et al 2015 [130]	Smartphone app	Automated and personalised suggestions and feedback on diet and PA using recommender systems
Verwey et al 2015 [131-133]	Smartphone app + Web	Annotate and comment on daily PA, share with HCP and goal-setting
Wadwha et al 2015 [134]	Smartphone app + web	Notifications to increase activity time, leadership board, monetary rewards
Wang et al 2015 [135]	Smartphone app + wearable + SMS	Activity prompt and goal-setting via SMS
Watson et al 2015 [136]	Wearable + wireless scales + web	Goal-setting and automated feedback on weight, diet and PA, meal suggestions, social support via community forum
Broekhuizen 2016 [137]	Wearable + web + SMS/email	Goal-setting, sustainability support, personal e-coach sends PA advice
Butryn et al 2016 [138]	Wearable + web	Goal-setting, leaderboard to support social comparison, social support
Choi et al 2016 [139]	Smartphone app + wearable	Daily text or video message, interactive automated feedback and advice, images and video clips regarding posture and stretching
Ciman et al 2016 [140]	Smartphone app	Stair-climbing game, with individual, collaborative and competitive modes
Darvall et al 2016 [141]	Wearable	Feedback on PA
Ding et al 2016 [142]	Smartphone app + wearable	Just-in-time context-aware reminders with explanations of context to support behaviour change
Fennell et al 2016 [143]	Wearable	Positive reinforcement through material incentives, negative reinforcement via monetary buy ins.
Garde et al 2016 [144]	Smartphone app + wearable	PA is rewarded by playtime and game incentives such as special features, unlocked levels, competitive score keeping, and peer interaction
Gilson et al 2016 [145]	Smartphone app + wearable	Feedback on PA, dietary logging
Glance et al 2016 [146]	Wearable + web	Team averages, reward badges, prizes, Leaderboards, Social Network
H-Jennings et al 2016 [147]	Wearable + web	Leaderboard to track individual daily progress, a forum for discussions, references on PA and sleep
Hartman et al 2016	Smartphone app + wearable + web	Self-regulatory skill building, goal-setting

[148]		
Herrmany et al 2016 [149]	Smartphone app	Automated goal recommendations, manual goal setting, record reference routes
Melton et al 2016 [150]	Smartphone app + wearable + web	Weekly health tips and reminders
Patel et al 2016 [151-153]	Smartphone app + wearable + SMS	Individual or combined financial incentives, different social comparison settings
Paul et al 2016 [154]	Smartphone app	Avatar, view when others' active, individual goal-setting, image-based rewards
Quintiliani et al 2016 [155]	Wearable + wireless scale + telephone	Telephone counselling with HCP
Vorrink et al 2016 [156]	Smartphone app + web + SMS	Website for physiotherapists to monitor patients and adjust goals
Walsh et al 2016 [157]	Smartphone app	Goal-setting, visually appealing graphic display of step-count history
Yingling et al 2016 [158]	Wearable + web	Online diary
Ashton et al 2017 [159]	Smartphone app + wearable + web	Educational website, social network
Chen et al 2017 [160]	Smartphone app + wearable	Team dyads, leaderboard and collective goal setting for social cooperation and comparison
Chung et al 2017 [161]	Smartphone app + wearable	Graphical feedback on PA, social network
Gell et al 2017 [162]	Wearable + SMS	Location-based feedback including a map of MVPA bout locations, discussion, goal-setting and planning with HCP
McMahon et al 2017 [163]	Wearable	Feedback on PA, small in-person group discussions to support interpersonal (social support, social comparison) and intrapersonal (goal-setting) behaviour change strategies
Neil-Sztramko et al 2017 [164]	Wearable + web	Feedback on PA, remote behavioural counselling sessions with HCP
Valle et al 2017 [165]	Wearable + wireless scales + web	Weight gain education, real-time feedback and weekly tailored feedback

a All interventions included sensor-based feedback on physical activity
b PA = Physical activity, HCP = Healthcare Professionals,

Appendix 3: Interview topic schedule for Study 2

Experiences

1. Would you mind introducing yourself? (***Prompt***: Name, where you work, position)
2. If academic: what discipline would you say you work within? If developer: can you tell me more about your company, and your **role** there?
3. Can you tell me a bit about the app/technologies you've evaluated and what you tend to evaluate / want to know?
    i. app's area of behavior change/purpose, technology/sensors used
    ii. what you want to find out/questions you have
    iii. **Did you evaluate behavior change – if so, why/why not?**
4. How did you evaluate these?
    i. Methods used: e.g. sensors, questionnaires, interviews....?
    ii. Experimentation: Did you use an experimental approach/ conduct any experiments/? If so, what do you test/ AND what was the set up? (***Prompt***: what research designs did you use, if any?)
5. How do you **choose** what methods and research designs to use / what **influences** your decision? **[*Prompt*:** why did you use X - time, resources...?]
6. What challenges did you encounter in evaluating the app(s)? (***Prompt*:** practical issues, working with others... anything else?)
7. How were challenges overcome, if at all?
8. Where do you get information about how to evaluate apps (what experimental designs and methods to use)?

Scoping Review questions

Research designs have been recommended specifically for health technologies (SCDs/N-of-1/ CEEBIT/MOST/SMART), because traditional methods (e.g. RCTs) don't always keep up with the pace of technology. Also been recommended that we use sensors to measure behaviour. We wanted to find out whether these experimental designs and methods are being used for physical activity apps and technologies.

9. We found not many researchers are using experimental designs recommended specifically for health technologies

i. What are your opinions on that (***Prompt***: **why** do you think that might be?) ***Prompt***: they were using traditional experimental designs instead (e.g. RCTs, pilots).
ii. How familiar are you with these recommended research designs (SCDs/N-of-1/ CEEBIT/MOST/SMART) (***Prompt***: have you heard of any of them? What are your experiences of using these?)
iii. What do you think are **benefits/values** of using these alternative designs, including single case designs, if any?
iv. What might stop you/others from using alternative designs? (***Prompt***: what are the risks, or downsides? [including single case designs?] (prompt practical, cultural (eg. funding), other)
v. We just studied physical activity apps and wearables – do you think it would be any different with the technologies you evaluate? (***Prompt*** – are new research designs used?)

10. The evaluations we were looking at were of apps and wearables that collected data to provide feedback on physical activity. Yet, we found some researchers are using additional, validated sensors to measure physical activity.
    i. What are your opinions on that – why do you think that might be?
    ii. How might this finding (using commercial vs validated/research grade sensors) apply to your area of research?
    iii. What are the benefits of using consumer-based sensors to measure behaviour, if any? (**Prompts:** scale, real-world representativeness, other)
    iv. What are the risks and barriers of using commercial devices to measure behaviour, if any? (**Prompts**: validity, industry reluctant to share algorithms)?

11. It was also of interest to find out how acceptability and usability were being evaluated. We found that when evaluating engagement / **use**, people mostly used log data, but acceptability or usability was mostly questionnaires or qualitative methods.
    i. What are your **opinions on using log data** to understand acceptability/usability? (**prompt**: benefits,

disadvantages/challenges/risks. Prompt – opinions on exploring acceptability/usability at **scale**/remotely?)

ii. How do you feel about launching large-scale **effectiveness** studies (using commercial/real-world platforms (such as the app store), without first testing for in small-scale local trials?

Last few questions

12. What do you feel is needed in order to progress and advance mobile health? / How are we doing ?
13. Overall, what, if any, kinds of support would be helpful for you, or others, in **evaluating behaviour change?** (e.g. training, access to literature, software, networking?)?

14. The idea is to eventually, in future, create evaluation guidance for developers so that they can log data that is of interest to behaviour change researchers, and even conduct experiments automatically using the app/app store.
    i. What issues/challenges might there be with this? What benefits might there be, if any?

List of References

Adcock, R. (2001). "Measurement validity: A shared standard for qualitative and quantitative research." *American political science review* **95**(3): 529-546.

Adler, M. and E. Ziglio (1996). *Gazing into the oracle: The Delphi method and its application to social policy and public health*, Jessica Kingsley Publishers.

Agboola, S., R. Palacholla, A. Centi, J. Kvedar and K. Jethwani (2016). "A Multimodal mHealth Intervention (FeatForward) to Improve Physical Activity Behavior in Patients with High Cardiometabolic Risk Factors: Rationale and Protocol for a Randomized Controlled Trial." *JMIR research protocols* **5**(2).

Aitken, M., B. Clancy and D. Nass The Growing Value of Digital Health Evidence and Impact on Human Health and the Healthcare System. Iqvia Institute; 7 November 2017.

Albaina, I. M., T. Visser, C. van der Mast and M. H. Vastenburg (2009). Flowie: A persuasive virtual coach to motivate elderly individuals to walk. 3rd International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2009).

Altrichter, H. and M. L. Holly (2005). "Research diaries." *Research methods in the social sciences*: 24-32.

Amorim, A. B., E. Pappas, M. Simic, M. L. Ferreira, A. Tiedemann, M. Jennings and P. H. Ferreira (2016). "Integrating Mobile health and Physical Activity to reduce the burden of Chronic low back pain Trial (IMPACT): a pilot trial protocol." *BMC Musculoskelet Disord* **17**: 36.

Anselm, S. and J. Corbin (1998). "Basics of qualitative research: Techniques and procedures for developing grounded theory." *Thousand Oaks, California: Saga Publication*.

Aranki, D., G. Kurillo, A. Mani, P. Azar, J. Van Gaalen, Q. Peng, P. Nigam, M. P. Reddy, S. Sankavaram and Q. Wu (2016). A telemonitoring framework for

android devices. Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016 IEEE First International Conference on, IEEE.

Arksey, H. and L. O'Malley (2005). "Scoping studies: towards a methodological framework." *International journal of social research methodology* **8**(1): 19-32.

Armstrong, R., B. J. Hall, J. Doyle and E. Waters (2011). "'Scoping the scope'of a cochrane review." *Journal of Public Health* **33**(1): 147-150.

Armstrong, T. and F. Bull (2006). "Development of the world health organization global physical activity questionnaire (GPAQ)." *Journal of Public Health* **14**(2): 66-70.

Arsand, E., N. Tatara, G. Ostengen and G. Hartvigsen (2010). "Mobile phone-based self-management tools for type 2 diabetes: the few touch application." *J Diabetes Sci Technol* **4**(2): 328-336.

Ashton, L. M., P. J. Morgan, M. J. Hutchesson, M. E. Rollo and C. E. Collins (2017). "Feasibility and preliminary efficacy of the 'HEYMAN'healthy lifestyle program for young men: a pilot randomised controlled trial." *Nutrition journal* **16**(1): 2.

Baer, D. M., M. M. Wolf and T. R. Risley (1968). "Some current dimensions of applied behavior analysis 1." *Journal of applied behavior analysis* **1**(1): 91-97.

Bajwa, S. S., X. Wang, A. N. Duc, R. M. Chanin, R. Prikladnicki, L. B. Pompermaier and P. Abrahamsson (2017). "Start-ups must be ready to pivot." *IEEE Software* **34**(3): 18-22.

Balas, E. A., & Boren, S. A. (2000). Managing clinical knowledge for health care improvement. Yearbook of medical informatics 2000: Patient-centered systems. Chicago

Barker, F., Atkins, L., & de Lusignan, S. (2016). Applying the COM-B behaviour model and behaviour change wheel to develop an intervention to improve

hearing-aid use in adult auditory rehabilitation. *International journal of audiology*, *55*(sup3), S90-S98.

Barriball, L.K. and While A. (1994). "Collecting Data using a semi-structured interview: a discussion paper." *Journal of advanced nursing* **19**(2): 328-335.

Barwais, F. A., T. F. Cuddihy and L. M. Tomson (2013). "Physical activity, sedentary behavior and total wellness changes among sedentary adults: a 4-week randomized controlled trial." *Health and quality of life outcomes* **11**(1): 183.

Barwais, F. A., T. F. Cuddihy, T. Washington, L. M. Tomson and E. Brymer (2014). "Development and validation of a new self-report instrument for measuring sedentary behaviors and light-intensity physical activity in adults." *Journal of Physical Activity and Health* **11**(6): 1097-1104.

Ben-Zeev, D., S. M. Schueller, M. Begale, J. Duffecy, J. M. Kane and D. C. Mohr (2015). "Strategies for mHealth research: Lessons from 3 mobile intervention studies." *Administration and Policy in Mental Health and Mental Health Services Research* **42**(2): 157-167.

Bentley, F., K. Tollmar, P. Stephenson, L. Levy, B. Jones, S. Robertson, E. Price, R. Catrambone and J. Wilson (2013). "Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change." *ACM Transactions on Computer-Human Interaction (TOCHI)* **20**(5): 30.

Besson, H., S. Brage, R. W. Jakes, U. Ekelund and N. J. Wareham (2010). "Estimating physical activity energy expenditure, sedentary time, and physical activity intensity by self-report in adults." *The American journal of clinical nutrition* **91**(1): 106-114.

Bhaskar, R. (1978). "A Realist Theory of Science, 2nd edn (Brighton, The Harvester Press)."

Bickmore, T. W., D. Mauer and T. Brown (2009). "Context awareness in a handheld exercise agent." *Pervasive and Mobile Computing* **5**(3): 226-235.

BinDhim, N. F., K. McGeechan and L. Trevena (2014). "Assessing the effect of an interactive decision-aid smartphone smoking cessation application (app) on quit rates: a double-blind automated randomised control trial protocol." *BMJ open* **4**(7): e005371.

Blackman, D. E. (2017). *Operant conditioning: an experimental analysis of behaviour*, Routledge.

Blackman, K. C. A., J. Zoellner, L. M. Berrey, R. Alexander, J. Fanning, J. L. Hill and P. A. Estabrooks (2013). "Assessing the Internal and External Validity of Mobile Health Physical Activity Promotion Interventions: A Systematic Literature Review Using the RE-AIM Framework." *Journal of Medical Internet Research* **15**(10): 81-95.

Blackwell, A. F. (2015). HCI as an Inter-Discipline. Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, ACM.

Blair, S. N. (2009). "Physical inactivity: the biggest public health problem of the 21st century." *British journal of sports medicine* **43**(1): 1-2.

Blandford, A., J. Gibbs, N. Newhouse, O. Perski, A. Singh and E. Murray (2018). "Seven lessons for interdisciplinary research on interactive digital health interventions." *Digital Health* **4**: 2055207618770325.

Bond, D. S., J. G. Thomas, H. A. Raynor, J. Moon, J. Sieling, J. Trautvetter, T. Leblond and R. R. Wing (2014). "B-MOBILE - A Smartphone-Based Intervention to Reduce Sedentary Time in Overweight/Obese Individuals: A Within-Subjects Experimental Trial." *Plos One* **9**(6).

Bondaronek, P., G. Alkhaldi, A. Slee, F. L. Hamilton and E. Murray (2018). "Quality of Publicly Available Physical Activity Apps: Review and Content Analysis." *JMIR Mhealth Uhealth* **6**(3): e53.

Borrelli, B. (2011). "The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials." *Journal of public health dentistry* **71**: S52-S63.

Bort-Roig, J., N. D. Gilson, A. Puig-Ribera, R. S. Contreras and S. G. Trost (2014). "Measuring and influencing physical activity with smartphone technology: a systematic review." *Sports Medicine* **44**(5): 671-686.

Bowen, D. J., M. Kreuter, B. Spring, L. Cofta-Woerpel, L. Linnan, D. Weiner, S. Bakken, C. P. Kaplan, L. Squiers and C. Fabrizio (2009). "How we design feasibility studies." *American journal of preventive medicine* **36**(5): 452-457.

Branch, J. L. (2000). "Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters." *Library & Information Science Research* **22**(4): 371-392.

Bravata, D. M., C. Smith-Spangler, V. Sundaram, A. L. Gienger, N. Lin, R. Lewis, C. D. Stave, I. Olkin and J. R. Sirard (2007). "Using pedometers to increase physical activity and improve health: a systematic review." *Jama* **298**(19): 2296-2304.

Brickwood, K.-J., S. T. Smith, G. Watson and A. D. Williams (2017). "The effect of ongoing feedback on physical activity levels following an exercise intervention in older adults: a randomised controlled trial protocol." *BMC Sports Science, Medicine and Rehabilitation* **9**(1): 1.

Broekhuizen, K., J. de Gelder, C. A. Wijsman, L. W. Wijsman, R. G. J. Westendorp, E. Verhagen, P. E. Slagboom, A. J. de Craen, W. van Mechelen, D. van Heemst, F. van der Ouderaa and S. P. Mooijaart (2016). "An Internet-Based Physical Activity Intervention to Improve Quality of Life of Inactive Older Adults: A Randomized Controlled Trial." *Journal of Medical Internet Research* **18**(4).

Brooke, J. (1996). "SUS-A quick and dirty usability scale." *Usability evaluation in industry* **189**(194): 4-7.

Brooke, S. M., H.-S. An, S.-K. Kang, J. M. Noble, K. E. Berg and J.-M. Lee (2017). "Concurrent validity of wearable activity trackers under free-living conditions." *Journal of strength and conditioning research* **31**(4): 1097-1106.

Brown, B., M. McGregor and E. Laurier (2013). iPhone in vivo: video analysis of mobile device use. Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM.

Bryman, A. and C. Cassell (2006). "The researcher interview: a reflexive perspective." *Qualitative Research in Organizations and Management: an international journal* **1**(1): 41-55.

Bulté, I. and P. Onghena (2012). "When the truth hits you between the eyes." *Methodology*.

Burns, P., C. Lueg and S. Berkovsky (2012). Activmon: encouraging physical activity through ambient social awareness. CHI'12 Extended Abstracts on Human Factors in Computing Systems, ACM.

Butryn, M. L., D. Arigo, G. A. Raggio, M. Colasanti and E. M. Forman (2016). "Enhancing physical activity promotion in midlife women with technology-based self-monitoring and social connectivity: A pilot study." *Journal of Health Psychology* **21**(8): 1548-1555.

Cadmus-Bertram, L. Marcus, R. E. Patterson, B. A. Parker and B. L. Morey (2015a). "Randomized trial of a fitbit-based physical activity intervention for women." *American Journal of Preventive Medicine* **49**(3): 414-418.

Cadmus-Bertram, L. Marcus, R. E. Patterson, B. A. Parker and B. L. Morey (2015b). "Use of the Fitbit to measure adherence to a physical activity intervention among overweight or obese, postmenopausal women: self-monitoring trajectory during 16 weeks." *JMIR mHealth and uHealth* **3**(4).

Carr, J. E. (2005). "Recommendations for reporting multiple-baseline designs across participants." *Behavioral Interventions: Theory & Practice in Residential & Community-Based Clinical Programs* **20**(3): 219-224.

Casey, M., P. S. Hayes, F. Glynn, G. ÓLaighin, D. Heaney, A. W. Murphy and L. G. Glynn (2014). "Patients' experiences of using a smartphone application to increase physical activity: the SMART MOVE qualitative study in primary care." *Br J Gen Pract* **64**(625): e500-e508.

Caulfield, B., I. Kaljo and S. Donnelly (2014). "Use of a consumer market activity monitoring and feedback device improves exercise capacity and activity levels in COPD." *2014 36th Annual International Conference of the Ieee Engineering in Medicine and Biology Society (Embc)*: 1765-1768.

Chatterjee, S., J. Byun, A. Pottathil, M. Moore, K. Dutta and H. Xie (2012). "Persuasive sensing: a novel in-home monitoring technology to assist elderly adult diabetic patients." *Persuasive Technology. Design for Health and Safety*: 31-42.

Chatterjee, S., K. Dutta, H. Xie, B. Jongbok, A. Pottathil and M. Moore (2013). Persuasive and pervasive sensing: A new frontier to monitor, track and assist older adults suffering from type-2 diabetes. System Sciences (HICSS), 2013 46th Hawaii International Conference on.

Chen, C., D. Haddad, J. Selsky, J. E. Hoffman, R. L. Kravitz, D. E. Estrin and I. Sim (2012). "Making sense of mobile health data: an open architecture to improve individual-and population-level health." *Journal of medical Internet research* **14**(4).

Chen, Y., Y. Chen, M. M. Randriambelonoro, A. Geissbuhler and P. Pu (2017). Design Considerations for Social Fitness Applications: Comparing Chronically Ill Patients and Healthy Adults. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland, Oregon, USA, ACM**:** 1753-1762.

Chen, Y. and P. Pu (2014). HealthyTogether: exploring social incentives for mobile fitness applications. Proceedings of the Second International Symposium of Chinese CHI, ACM.

Chimen, M., A. Kennedy, K. Nirantharakumar, T. Pang, R. Andrews and P. Narendran (2012). "What are the health benefits of physical activity in type 1 diabetes mellitus? A literature review." *Diabetologia* **55**(3): 542-551.

Choi, J., J. hyeon Lee, E. Vittinghoff and Y. Fukuoka (2016). "mHealth physical activity intervention: a randomized pilot study in physically inactive pregnant women." *Maternal and child health journal* **20**(5): 1091-1101.

Chorpita, B. F., E. L. Daleiden and J. R. Weisz (2005). "Modularity in the design and application of therapeutic interventions." *Applied and Preventive Psychology* **11**(3): 141-156.

Christ, T. J. (2007). "Experimental control and threats to internal validity of concurrent and nonconcurrent multiple baseline designs." *Psychology in the Schools* **44**(5): 451-459.

Chung, A. E., A. C. Skinner, S. E. Hasty and E. M. Perrin (2017). "Tweeting to health: a novel mHealth intervention using Fitbits and Twitter to foster healthy lifestyles." *Clinical Pediatrics* **56**(1): 26-32.

Church, T. S., Y. J. Cheng, C. P. Earnest, C. E. Barlow, L. W. Gibbons, E. L. Priest and S. N. Blair (2004). "Exercise capacity and body composition as predictors of mortality among men with diabetes." *Diabetes care* **27**(1): 83-88.

Ciman, M., M. Donini, O. Gaggi and F. Aiolli (2016). "Stairstep recognition and counting in a serious Game for increasing users' physical activity." *Personal and Ubiquitous Computing* **20**(6): 1015-1033.

Clayton, C., L. Feehan, C. H. Goldsmith, W. C. Miller, N. Grewal, J. Ye, J. Y. Yoo and L. C. Li (2015). "Feasibility and preliminary efficacy of a physical activity counseling intervention using Fitbit in people with knee osteoarthritis: the TRACK-OA study protocol." *Pilot and Feasibility Studies* **1**(1): 30.

Coiera E. Four rules for the reinvention of health care. BMJ. 2004 May 15;328(7449):1197–9. doi: 10.1136/bmj.328.7449.1197.

Collins, L. M., S. A. Murphy, V. N. Nair and V. J. Strecher (2005). "A strategy for optimizing and evaluating behavioral interventions." *Annals of Behavioral Medicine* **30**(1): 65-73.

Collins, L. M., S. A. Murphy and V. Strecher (2007). "The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions." *American journal of preventive medicine* **32**(5): S112-S118.

Concato, J., N. Shah and R. I. Horwitz (2000). "Randomized, controlled trials, observational studies, and the hierarchy of research designs." *New England Journal of Medicine* **342**(25): 1887-1892.

Connelly, K. (2007). On developing a technology acceptance model for pervasive computing. 9th International Conference on Ubiquitous Computing (UBICOMP)-Workshop of Ubiquitous System Evaluation (USE), Springer, Innsbruck, Austria.

Consolvo, S., P. Klasnja, D. W. McDonald, D. Avrahami, J. Froehlich, L. LeGrand, R. Libby, K. Mosher and J. A. Landay (2008). Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. Proceedings of the 10th international conference on Ubiquitous computing, ACM.

Cook, T. D., D. T. Campbell and A. Day (1979). *Quasi-experimentation: Design & analysis issues for field settings*, Houghton Mifflin Boston.

Cooper, A. J., K. Dearnley, K. M. Williams, S. J. Sharp, E. M. van Sluijs, S. Brage, S. Sutton and S. J. Griffin (2015). "Protocol for Get Moving: a randomised controlled trial to assess the effectiveness of three minimal contact interventions to promote fitness and physical activity in working adults." *BMC public health* **15**(1): 296.

Cornelius, C. and D. Kotz (2010). On usable authentication for wireless body area networks. Proceedings of the First USENIX Workshop on Health Security and Privacy (HealthSec).

Costa, C. E. and C. R. X. Cançado (2012). "Stability check: A program for calculating the stability of behavior." *Revista Mexicana de Análisis de la Conducta* **38**(1).

Cowan, L. T., S. A. Van Wagenen, B. A. Brown, R. J. Hedin, Y. Seino-Stephan, P. C. Hall and J. H. West (2013). "Apps of steel: are exercise apps providing consumers with realistic expectations? A content analysis of exercise apps for presence of behavior change theory." *Health Education & Behavior* **40**(2): 133-139.

Craig, C. L., A. L. Marshall, M. Sjöström, A. E. Bauman, M. L. Booth, B. E. Ainsworth, M. Pratt, U. Ekelund, A. Yngve and J. F. Sallis (2003). "International physical activity questionnaire: 12-country reliability and validity." *Medicine & Science in Sports & Exercise* **35**(8): 1381-1395.

Craig, P., P. Dieppe, S. Macintyre, S. Michie, I. Nazareth and M. Petticrew (2008). "Developing and evaluating complex interventions: the new Medical Research Council guidance." *Bmj* **337**: a1655.

Crane, D., C. Garnett, S. Michie, R. West and J. Brown (2018). "A smartphone app to reduce excessive alcohol consumption: Identifying the effectiveness of intervention components in a factorial randomised control trial." *Scientific reports* **8**(1): 4384.

Cresswell, K. M., A. Blandford and A. Sheikh (2017). "Drawing on human factors engineering to evaluate the effectiveness of health information technology." *Journal of the Royal Society of Medicine* **110**(8): 309-315.

Creswell, J. W. and V. L. P. Clark (2007). "Designing and conducting mixed methods research."

Creswell, J. W., A. C. Klassen, V. L. Plano Clark and K. C. Smith (2011). "Best practices for mixed methods research in the health sciences." *Bethesda (Maryland): National Institutes of Health*: 2094-2103.

Creswell, J. W. and D. L. Miller (2000). "Determining validity in qualitative inquiry." *Theory into practice* **39**(3): 124-130.

Curran, G. M., M. Bauer, B. Mittman, J. M. Pyne and C. Stetler (2012). "Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact." *Medical care* **50**(3): 217.

d'Heureuse, N., F. Huici, M. Arumaithurai, M. Ahmed, K. Papagiannaki and S. Niccolini (2012). "What's app?: a wide-scale measurement study of smart phone markets." *ACM SIGMOBILE Mobile Computing and Communications Review* **16**(2): 16-27.

Dallery, J., R. N. Cassidy and B. R. Raiff (2013). "Single-case experimental designs to evaluate novel technology-based health interventions." *Journal of medical Internet research* **15**(2): e22.

Dallery, J. and B. Raiff (2014). "Optimizing behavioral health interventions with single-case designs: from development to dissemination." *Translational Behavioral Medicine* **4**(3): 290-303.

Darvall, J., A. Parker and D. Story (2016). "Feasibility and acceptability of remotely monitored pedometer-guided physical activity." *Anaesthesia and intensive care* **44**(4): 501.

Daskalova, N., D. Metaxa-Kakavouli, A. Tran, N. Nugent, J. Boergers, J. McGeary and J. Huang (2016). SleepCoacher: A personalized automated self-experimentation system for sleep recommendations. Proceedings of the 29th Annual Symposium on User Interface Software and Technology, ACM.

Davies, E. B., M. P. Craven, J. L. Martin and L. Simons (2017). "Proportionate methods for evaluating a simple digital mental health tool." *Evidence-based mental health*: ebmental-2017-102755.

Davis, H. O. (2001). System and method for conducting focus groups using remotely loaded participants over a computer network, Google Patents.

Deci, E. L. and R. M. Ryan (2003). "Intrinsic motivation inventory." *Self-Determination Theory* **267**.

Denzin, N. (1970). "Strategies of multiple triangulation." *The research act in sociology: A theoretical introduction to sociological method* **297**: 313.

Ding, X., J. Xu, H. H. Wang, G. L. Chen, H. Thind, and Y. Zhang (2016). "WalkMore: Promoting Walking with Just-in-Time Context-Aware Prompts." *2016 IEEE Wireless Health (Wh)*: 65-72.

Dipietro, L., C. J. Caspersen, A. M. Ostfeld and E. R. Nadel (1993). "A survey for assessing physical activity among older adults." *Medicine & Science in Sports & Exercise*.

Direito, A., E. Carraça, J. Rawstorn, R. Whittaker and R. Maddison (2016). "mHealth technologies to influence physical activity and sedentary behaviors: behavior change techniques, systematic review and meta-analysis of randomized controlled trials." *Annals of behavioral medicine* **51**(2): 226-239.

Direito, A., L. P. Dale, E. Shields, R. Dobson, R. Whittaker and R. Maddison (2014). "Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques?" *Bmc Public Health* **14**.

Direito, A., Y. N. Jiang, R. Whittaker and R. Maddison (2015). "Apps for IMproving FITness and Increasing Physical Activity Among Young People: The AIMFIT Pragmatic Randomized Controlled Trial." *Journal of Medical Internet Research* **17**(8).

Dobkin, B. H. and A. Dorsch (2011). "The Promise of mHealth: Daily Activity Monitoring and Outcome Assessments by Wearable Sensors." *Neurorehabilitation and Neural Repair* **25**(9): 788-798.

Dobkin, B. H. and C. Martinez (2018). "Wearable Sensors to Monitor, Enable Feedback, and Measure Outcomes of Activity and Practice." *Current neurology and neuroscience reports* **18**(12): 87.

Donkin, L., H. Christensen, S. L. Naismith, B. Neal, I. B. Hickie and N. Glozier (2011). "A systematic review of the impact of adherence on the effectiveness of e-therapies." *Journal of medical Internet research* **13**(3): e52.

Dumais, S., R. Jeffries, D. M. Russell, D. Tang and J. Teevan (2014). *Understanding user behavior through log data and analysis*. *Ways of Knowing in HCI*, Springer**:** 349-372.

Duncan, M. J., C. Vandelanotte, S. G. Trost, A. L. Rebar, N. Rogers, N. W. Burton, B. Murawski, A. Rayward, S. Fenton and W. J. Brown (2016). "Balanced: a randomised trial examining the efficacy of two self-monitoring methods for an app-based multi-behaviour intervention to improve physical activity, sitting and sleep in adults." *Bmc Public Health* **16**.

Duncan, M. J., K. Wunderlich, Y. Zhao and G. Faulkner (2018). "Walk this way: validity evidence of iphone health application step count in laboratory and free-living conditions." *Journal of Sports Sciences* **36**(15): 1695-1704.

Dunton, G. F., E. Dzubur, K. Kawabata, B. Yanez, B. Bo and S. Intille (2014). "Development of a smartphone application to measure physical activity using sensor-assisted self-report." *Front Public Health* **2**: 12.

Dzewaltowski, D. A., P. A. Estabrooks and R. E. Glasgow (2004). "The future of physical activity behavior change research: what is needed to improve translation of research into health promotion practice?" *Exercise and sport sciences reviews* **32**(2): 57-63.

El-Nasr, M. S., A. Drachen and A. Canossa (2013). "Game analytics." *New York, Sprint*.

Elo, S. and H. Kyngäs (2008). "The qualitative content analysis process." *Journal of advanced nursing* **62**(1): 107-115.

Eng, D. S. and J. M. Lee (2013). "The promise and peril of mobile health applications for diabetes and endocrinology." *Pediatric diabetes* **14**(4): 231-238.

Evans, D. and A. Pearson (2001). "Systematic reviews: gatekeepers of nursing knowledge." *Journal of Clinical Nursing* **10**(5): 593-599.

Eysenbach, G. (2005). The law of attrition. *Journal of medical Internet research*, *7*(1).

Eysenbach, G. and CONSORT E-Health Group (2011). "CONSORT-EHEALTH: Improving and Standardizing Evaluation Reports of Web-based and Mobile Health Interventions." *Journal of Medical Internet Research* **13**(4): e126.

Faridi, Z., L. Liberti, K. Shuval, V. Northrup, A. Ali and D. L. Katz (2008). "Evaluating the impact of mobile telephone technology on type 2 diabetic patients' self-management: The NICHE pilot study." *Journal of Evaluation in Clinical Practice* **14**(3): 465-469.

Fennell, C., H. Gerhart, Y. Seo, K. Hauge and E. L. Glickman (2016). "Combined incentives versus no-incentive exercise programs on objectively measured physical activity and health-related variables." *Physiology & behavior* **163**: 245-250.

Ferguson, T., A. V. Rowlands, T. Olds and C. Maher (2015). "The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study." *International Journal of Behavioral Nutrition and Physical Activity* **12**(1): 42.

Ferrari, R. (2015). "Writing narrative style literature reviews." *Medical Writing* **24**(4): 230-235.

Ferreira, D., Kostakos, V., & Dey, A. K. (2012). Lessons learned from large-scale user studies: Using android market as a source of data. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, *4*(3), 28-43.

Fialho, A. T., H. van den Heuvel, Q. Shahab, Q. Liu, L. Li, P. Saini, J. Lacroix and P. Markopoulos (2009). ActiveShare: sharing challenges to increase physical activities. CHI'09 Extended Abstracts on Human Factors in Computing Systems, ACM.

Finkelstein, E. A., B. A. Haaland, M. Bilger, A. Sahasranaman, R. A. Sloan, E. E. K. Nang and K. R. Evenson (2016). "Effectiveness of activity trackers with and without incentives to increase physical activity (TRIPPA): a randomised controlled trial." *Lancet Diabetes & Endocrinology* **4**(12): 983-995.

Finkelstein, E. A., A. Sahasranaman, G. John, B. A. Haaland, M. Bilger, R. A. Sloan, E. E. K. Nang and K. R. Evenson (2015). "Design and baseline characteristics of participants in the TRial of Economic Incentives to Promote Physical Activity (TRIPPA): A randomized controlled trial of a six month pedometer program with financial incentives." *Contemporary Clinical Trials* **41**: 238-247.

Fisher, W. W., M. E. Kelley and J. E. Lomas (2003). "Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs." *Journal of applied behavior analysis* **36**(3): 387-406.

Fitzsimons, C. F., A. Kirk, G. Baker, F. Michie, C. Kane and N. Mutrie (2013). "Using an individualised consultation and activPAL™ feedback to reduce sedentary time in older Scottish adults: results of a feasibility and pilot study." *Preventive medicine* **57**(5): 718-720.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive medicine*, *15*(5), 451-474.

Frederix, I., D. Hansen, K. Coninx, P. Vandervoort, D. Vandijck, N. Hens, E. Van Craenenbroeck, N. Van Driessche and P. Dendale (2015). "Medium-term effectiveness of a comprehensive Internet-based and patient-specific telerehabilitation program with text messaging support for cardiac patients: randomized controlled trial." *Journal of medical Internet research* **17**(7): e185.

Frederix, I., N. Van Driessche, D. Hansen, J. Berger, K. Bonne, T. Alders and P. Dendale (2015). "Increasing the medium-term clinical benefits of hospital-based cardiac rehabilitation by physical activity telemonitoring in coronary artery disease patients." *European Journal of Preventive Cardiology* **22**(2): 150-158.

Fujiki, Y., K. Kazakos, C. Puri, P. Buddharaju, I. Pavlidis and J. Levine (2008). "NEAT-o-Games: blending physical activity and fun in the daily routine." *Computers in Entertainment (CIE)* **6**(2): 21.

Fujiki, Y., K. Kazakos, C. Puri, I. Pavlidis, J. Starren and J. Levine (2007). NEAT-o-games: ubiquitous activity-based gaming. CHI'07 extended abstracts on Human factors in computing systems, ACM.

Gale, N. K., G. Heath, E. Cameron, S. Rashid and S. Redwood (2013). "Using the framework method for the analysis of qualitative data in multi-disciplinary health research." *BMC medical research methodology* **13**(1): 117.

Garde, A., A. Umedaly, S. M. Abulnaga, A. Junker, J. P. Chanoine, M. Johnson, J. M. Ansermino and G. A. Dumont (2016). "Evaluation of a Novel Mobile Exergame in a School-Based Environment." *Cyberpsychology, Behavior, and Social Networking* **19**(3): 186-192.

Garde, A., A. Umedaly, S. M. Abulnaga, L. Robertson, A. Junker, J. P. Chanoine, J. M. Ansermino and G. A. Dumont (2015). "Assessment of a mobile game ('MobileKids Monster Manor') to promote physical activity among children." *Games for Health* **4**(2): 145-148.

Gast, D. L. and A. D. Spriggs (2010). "Visual analysis of graphic data." *Single subject research methodology in behavioral sciences*: 199-233.

Gell, N. M., K. W. Grover, M. Humble, M. Sexton and K. Dittus (2017). "Efficacy, feasibility, and acceptability of a novel technology-based intervention to support physical activity in cancer survivors." *Supportive Care in Cancer* **25**(4): 1291-1300.

Geraedts, H. A., W. Zijlstra, W. Zhang, S. Bulstra and M. Stevens (2014). "Adherence to and effectiveness of an individually tailored home-based exercise program for frail older adults, driven by mobility monitoring: design of a prospective cohort study." *BMC Public Health* **14**: 570.

Giardino, C., M. Unterkalmsteiner, N. Paternoster, T. Gorschek and P. Abrahamsson (2014). "What do we know about software development in startups?" *IEEE software* **31**(5): 28-32.

Gilliland, J., R. Sadler, A. Clark, C. O'Connor, M. Milczarek and S. Doherty (2015). "Using a smartphone application to promote healthy dietary behaviours and local food consumption." *BioMed research international* **2015**.

Gilson, N. D., T. G. Pavey, C. Vandelanotte, M. J. Duncan, S. R. Gomersall, S. G. Trost and W. J. Brown (2015). "Chronic disease risks and use of a smartphone application during a physical activity and dietary intervention in Australian truck drivers." *Australian and New Zealand journal of public health*.

Glance, D. G., E. Ooi, Y. Berman, C. F. Glance, P. H. and R. Barrett (2016). "Impact of a Digital Activity Tracker-based Workplace Activity Program on Health and Wellbeing." *Dh'16: Proceedings of the 2016 Digital Health Conference*: 37-41.

Glaser, B. G., and Strauss, A. L., (1999). The Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Aldine De Gruyter.

Glasgow, R. E., E. Lichtenstein and A. C. Marcus (2003). "Why Don't We See More Translation of Health Promotion Research to Practice? Rethinking the Efficacy-to-Effectiveness Transition." *American Journal of Public Health* **93**(8): 1261-1267.

Glynn, L. G., P. S. Hayes, M. Casey, F. Glynn, A. Alvarez-Iglesias, J. Newell, G. Olaighin, D. Heaney and A. W. Murphy (2013). "SMART MOVE - a smartphone-based intervention to promote physical activity in primary care: study protocol for a randomized controlled trial." *Trials* **14**.

Glynn, L. G., P. S. Hayes, M. Casey, F. Glynn, A. Alvarez-Iglesias, J. Newell, G. Olaighin, D. Heaney, M. O'Donnell and A. W. Murphy (2014). "Effectiveness of a smartphone application to promote physical activity in primary care: the SMART MOVE randomised controlled trial." *British Journal of General Practice* **64**(624): E384-E391.

Godin, G. and R. Shephard (1985). "A simple method to assess exercise behavior in the community." *Can J Appl Sport Sci* **10**(3): 141-146.

Gomes, N., D. Merugu, G. O'Brien, C. Mandayam, Y. Jia Shuo, B. Atikoglu, A. Albert, N. Fukumoto, L. Huan, B. Prabhakar and D. Wischik (2012). Steptacular: An incentive mechanism for promoting wellness. Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on.

Goris, A. and R. Holmes (2008). The effect of a lifestyle activity intervention program on improving physical activity behavior of employees. Proceedings of the Third International Conference on Persuasive Technology.

Gouveia, R., E. Karapanos and M. Hassenzahl (2015). How do we engage with activity trackers?: a longitudinal study of Habito. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM.

Granado-Font, E., G. Flores-Mateo, M. Sorli-Aguilar, X. Montana-Carreras, C. Ferre-Grau, M. L. Barrera-Uriarte, E. Oriol-Colominas, C. Rey-Renones, I. Caules, E. M. Satue-Gracia and O. S. Grp (2015). "Effectiveness of a Smartphone application and wearable device for weight loss in overweight or obese primary care patients: protocol for a randomised controlled trial." *Bmc Public Health* **15**.

Grant, A., S. Treweek, T. Dreischulte, R. Foy and B. Guthrie (2013). "Process evaluations for cluster-randomised trials of complex interventions: a proposed framework for design and reporting." *Trials* **14**(1): 15.

Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. BMJ. 2010;341:c5814.]

Oxford Centre for Evidence-Based Medicine (2011): The Oxford levels of evidence.

Guthrie, N., A. Bradlyn, S. K. Thompson, S. Yen, J. Haritatos, F. Dillon and S. W. Cole (2015). "Development of an Accelerometer-Linked Online Intervention System to Promote Physical Activity in Adolescents." *Plos One* **10**(5).

Guyatt, G. H., R. B. Haynes, R. Z. Jaeschke, D. J. Cook, L. Green, C. D. Naylor, M. C. Wilson, W. S. Richardson and E.-B. M. W. Group (2000). "Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care." *Jama* **284**(10): 1290-1296.

Guyatt, G. H., J. L. Keller, R. Jaeschke, D. Rosenbloom, J. D. Adachi and M. T. Newhouse (1990). "The n-of-1 randomized controlled trial: clinical usefulness: our three-year experience." *Annals of internal medicine* **112**(4): 293-299.

H-Jennings, F., M.-V. Clément, M. Brown, B. Leong, L. Shen and C. Dong (2016). "Promote Students' Healthy Behavior Through Sensor and Game: A Randomized Controlled Trial." *Medical Science Educator* **26**(3): 349-355.

Hammersley, M. Atkinson, p.(1995) Ethnography: Principles in practice, London: Routledge.

Harman, M., Y. Jia and Y. Zhang (2012). App store mining and analysis: MSR for app stores. Proceedings of the 9th IEEE Working Conference on Mining Software Repositories, IEEE Press.

Harries, T., P. Eslambolchilar, R. Rettie, C. Stride, S. Walton and H. C. van Woerden (2016). "Effectiveness of a smartphone app in increasing physical activity amongst male adults: a randomised controlled trial." *Bmc Public Health* **16**.

Harries, T., P. Eslambolchilar, C. Stride, R. Rettie and S. Walton (2013). Walking in the wild–Using an always-on smartphone application to increase physical activity. IFIP Conference on Human-Computer Interaction, Springer.

Hartman, S. J., S. H. Nelson, L. A. Cadmus-Bertram, R. E. Patterson, B. A. Parker and J. P. Pierce (2016). "Technology-and Phone-Based Weight Loss Intervention: Pilot RCT in Women at Elevated Breast Cancer Risk." *American Journal of Preventive Medicine* **51**(5): 714-721.

Hartmann, D. P. and R. V. Hall (1976). "The changing criterion design." *Journal of Applied Behavior Analysis* **9**(4): 527-532.

Hartwig, M. (2015). *Dictionary of critical realism*, Routledge.

Haskell, W. L., I.-M. Lee, R. R. Pate, K. E. Powell, S. N. Blair, B. A. Franklin, C. A. Macera, G. W. Heath, P. D. Thompson and A. Bauman (2007). "Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association." *Circulation* **116**(9): 1081.

Hatcher, A. M., & Bonell, C. P. (2016). High time to unpack the 'how'and 'why' of adherence interventions. *Aids*, *30*(8), 1301-1303.

Haus, E., A. Reinberg, B. Mauvieux, N. Le Floc'h, L. Sackett-Lundeen and Y. Touitou (2016). "Risk of obesity in male shift workers: A chronophysiological approach." *Chronobiology International* **33**(8): 1018-1036.

Australian Institute of Health and Welfare AIHW (2003). The Active Australia Survey: A guide and manual for implementation, analysis and reporting [accessed 3/01/2018] www.aihw.gov.au/getmedia/ff25c134-5df2-45ba-b4e1-6c214ed157e6/aas.pdf.aspx?inline=true

Hekler, E. B., P. Klasnja, J. E. Froehlich and M. P. Buman (2013). Mind the theoretical gap: interpreting, using, and developing behavioral theory in HCI research. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM.

Hekler, E. B., P. Klasnja, W. T. Riley, M. P. Buman, J. Huberty, D. E. Rivera and C. A. Martin (2016). "Agile science: creating useful products for behavior change in the real world." *Translational Behavioral Medicine* **6**(2): 317-328.

Henderson, J., J. Noell, T. Reeves, T. Robinson and V. Strecher (1999). "Developers and evaluation of interactive health communication applications1." *American journal of preventive medicine* **16**(1): 30-34.

Hennink, M. M. (2013). *Focus group discussions*, Oxford University Press.

Henze, N. and S. Boll (2010). Push the study to the app store: Evaluating off-screen visualizations for maps in the android market. Proceedings of the 12th international conference on Human computer interaction with mobile devices and services, ACM.

Herrmanny, K., J. Ziegler and A. Dogangün (2016). *Supporting Users in Setting Effective Goals in Activity Tracking*. *Persuasive Technology: 11th International Conference, PERSUASIVE 2016, Salzburg, Austria, April 5-7, 2016, Proceedings*. A. Meschtscherjakov, B. De Ruyter, V. Fuchsberger, M. Murer and M. Tscheligi. Cham, Springer International Publishing**:** 15-26.

Hickey, A. M. and P. S. Freedson (2016). "Utility of consumer physical activity trackers as an intervention tool in cardiovascular disease prevention and treatment." *Progress in cardiovascular diseases* **58**(6): 613-619.

Hill, J. O. and H. R. Wyatt (2005). "Role of physical activity in preventing and treating obesity." *Journal of Applied Physiology* **99**(2): 765-770.

Hirano, S. H., R. G. Farrell, C. M. Danis and W. A. Kellogg (2013). WalkMinder: encouraging an active lifestyle using mobile phone interruptions. CHI'13 Extended Abstracts on Human Factors in Computing Systems, ACM.

Horner, R. H., E. G. Carr, J. Halle, G. McGee, S. Odom and M. Wolery (2005). "The use of single-subject research to identify evidence-based practice in special education." *Exceptional children* **71**(2): 165-179.

Horvath, A. O. and L. S. Greenberg (1989). "Development and validation of the Working Alliance Inventory." *Journal of counseling psychology* **36**(2): 223.

Hurley, J. C., K. E. Hollingshead, M. Todd, C. L. Jarrett, W. J. Tucker, S. S. Angadi and M. A. Adams (2015). "The Walking Interventions Through Texting (WalkIT) Trial: Rationale, Design, and Protocol for a Factorial Randomized Controlled Trial of Adaptive Interventions for Overweight and Obese, Inactive Adults." *JMIR Research Protocols* **4**(3): e108.

Hurling, R., M. Catt, M. D. Boni, B. W. Fairley, T. Hurst, P. Murray, A. Richardson and J. S. Sodhi (2007). "Using internet and mobile phone technology to deliver an automated physical activity program: Randomized controlled trial." *Journal of Medical Internet Research* **9**(2).

IMS Institute for Healthcare Informatics. IMS Health. 2015. Patient Adoption of mHealth URL: http://www.imshealth.com/ en/thought-leadership/ims-institute/reports/patient-adoption-of-mhealth [accessed 28/10/18]

Ioannidis JP: Effect of the statistical significance of results on time to completion and publication of randomized efficacy trials. JAMA 1998, 279:281–286.

Ioannidis, J. P., S. Greenland, M. A. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F. Schulz and R. Tibshirani (2014). "Increasing value and reducing waste in research design, conduct, and analysis." *The Lancet* **383**(9912): 166-175.

Iribarren, S. J., K. Cato, L. Falzon and P. W. Stone (2017). "What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions." *PLOS ONE* **12**(2): e0170581.

International Organization for Standardization (ISO) (1998). "9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)." *The international organization for standardization* **45**.

Israel, S. A., J. M. Irvine, A. Cheng, M. D. Wiederhold and B. K. Wiederhold (2005). "ECG to identify individuals." *Pattern recognition* **38**(1): 133-142.

Jahns, R. and P. Houck (2013). "Mobile Health Market Report 2013-2017." *Research2Guidance*.

Jake-Schoffman, D. E., V. J. Silfee, M. E. Waring, E. D. Boudreaux, R. S. Sadasivam, S. P. Mullen, J. L. Carey, R. B. Hayes, E. Y. Ding and G. G. Bennett (2017). "Methods for Evaluating the Content, Usability, and Efficacy of Commercial Mobile Health Apps." *JMIR mHealth and uHealth* **5**(12).

Jalali, S. and C. Wohlin (2012). Systematic literature studies: database searches vs. backward snowballing. Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement, ACM.

Jimenez Garcia, J., N. A. Romero, D. Keyson and P. Havinga (2013). ESTHER 1.3: integrating in-situ prompts to trigger self-reflection of physical activity in knowledge workers. Proceedings of the 2013 Chilean Conference on Human-Computer Interaction, ACM.

Johnson, R. B. (1997). "Examining the validity structure of qualitative research." *Education* **118**(2): 282.

Johnston, D. W. and M. Johnston (2013). "Useful theories should apply to individuals." *British journal of health psychology* **18**(3): 469-473.

Jones, D., N. Skrepnik, R. M. Toselli and B. Leroy (2016). "Incorporating Novel Mobile Health Technologies Into Management of Knee Osteoarthritis in Patients Treated With Intra-Articular Hyaluronic Acid: Rationale and Protocol of a Randomized Controlled Trial." *JMIR Research Protocols* **5**(3).

Karkar, R., J. Zia, R. Vilardaga, S. R. Mishra, J. Fogarty, S. A. Munson and J. A. Kientz (2015). "A framework for self-experimentation in personalized health." *Journal of the American Medical Informatics Association* **23**(3): 440-448.

Kay, M., G. L. Nelson and E. B. Hekler (2016). Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings*, Oxford University Press.

Kazdin, A. E. and S. A. Kopel (1975). "On resolving ambiguities of the multiple-baseline design: Problems and recommendations." *Behavior Therapy* **6**(5): 601-608.

Kennedy, M. M. (1979). "Generalizing from single case studies." *Evaluation quarterly* **3**(4): 661-678.

Khalil, A. and S. Abdallah (2013). "Harnessing social dynamics through persuasive technology to promote healthier lifestyle." *Computers in Human Behavior* **29**(6): 2674-2681.

Khan, A. M., S. W. Lee and Ieee (2013). "Need for a Context-Aware Personalized Health Intervention System to Ensure Long-Term Behavior Change to Prevent Obesity." *2013 5th International Workshop on Software Engineering in Health Care (Sehc)*: 71-74.

Kharrazi, H. and L. Vincz (2011). "Increasing physical activity by implementing a behavioral change intervention using pervasive personal health record system: an exploratory study." *Universal Access in Human-Computer Interaction. Applications and Services*: 366-375.

Killeen, P. R. (2018). "Predict, Control, and Replicate to Understand: How Statistics Can Foster the Fundamental Goals of Science." *Perspectives on Behavior Science*: 1-24.

King, A. C., E. B. Hekler, L. A. Grieco, S. J. Winter, J. L. Sheats, M. P. Buman, B. Banerjee, T. N. Robinson and J. Cirimele (2013). "Harnessing Different Motivational Frames via Mobile Phones to Promote Daily Physical Activity and Reduce Sedentary Behavior in Aging Adults." *Plos One* **8**(4).

King, A. C., E. B. Hekler, L. A. Grieco, S. J. Winter, J. L. Sheats, M. P. Buman, B. Banerjee, T. N. Robinson and J. Cirimele (2016). "Effects of Three Motivationally Targeted Mobile Device Applications on Initial Physical Activity and Sedentary Behavior Change in Midlife and Older Adults: A Randomized Trial." *Plos One* **11**(6).

King, R., Churchill, E.F., & Tan, C. (2017). *Designing with data: Improving the user experience with A/B testing*, O'Reilly Media.

Kirkevold, M. (1997). Integrative nursing research—an important strategy to further the development of nursing science and nursing practice. *Journal of advanced nursing*, *25*(5), 977-984.

Kitzinger, J. (1995). "Qualitative research: introducing focus groups." *Bmj* **311**(7000): 299-302.

Klasnja, P., S. Consolvo and W. Pratt (2011). How to evaluate technologies for health behavior change in HCI research. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM.

Klasnja, P., S. Consolvo and W. Pratt (2011). How to evaluate technologies for health behavior change in HCI research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Vancouver, BC, Canada, ACM**:** 3063-3072.

Klein, L. A., D. Houlihan, J. L. Vincent and C. J. Panahon (2017). "Best Practices in Utilizing the Changing Criterion Design." *Behavior Analysis in Practice* **10**(1): 52-61.

Knowler, W., E. Barrett-Connor, S. Fowler, F. Richard, R. Hamman, J. Lachin, E. Walker and D. Nathan (2002). "for the Diabetes Prevention Program Research Group the Diabetes Prevention Program Research Group: Reduction in the incidence of Type 2 diabetes with lifestyle intervention or metformin." *New Engl J Med* **346**: 393-403.

Komninos, A., M. D. Dunlop, D. Rowe, A. Hewitt, S. Coull and Ieee (2015). *Using Degraded Music Quality to Encourage a Health Improving Walking Pace: BeatClearWalker*. *Proceedings of the 2015 9th International Conference on Pervasive Computing Technologies for Healthcare***:** 57-64.

Kranz, M., L. Murmann and F. Michahelles (2013). "Research in the large: Challenges for large-scale mobile application research-a case study about NFC adoption using gamification via an app store." *International Journal of Mobile Human Computer Interaction (IJMHCI)* **5**(1): 45-61.

Kratochwill, T. R. (2014). Task force on evidence-based interventions in school psychology, Retrieved September 30th.

Kratochwill, T. R., J. Hitchcock, R. Horner, J. R. Levin, S. Odom, D. Rindskopf and W. Shadish (2010). "Single-case designs technical documentation." *What works clearinghouse*.

Kratochwill, T. R. and J. R. Levin (2010). "Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue." *Psychological Methods* **15**(2): 124.

Kravitz, R. (2016) and the DEcIDE Methods Center N-of-1 Guidance Panel (Duan N, Eslick I, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S). Design and implementation of N-of-1 trials: a user's guide. AHRQ Publication No. 13 (14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality; January 2014.

Kravitz, R. L., Paterniti, D. A., Hay, M. C., Subramanian, S., Dean, D. E., Weisner, T., ... & Duan, N. (2009). Marketing therapeutic precision: potential facilitators and barriers to adoption of n-of-1 trials. *Contemporary clinical trials*, *30*(5), 436-445.

Kumar, S., W. J. Nilsen, A. Abernethy, A. Atienza, K. Patrick, M. Pavel, W. T. Riley, A. Shar, B. Spring and D. Spruijt-Metz (2013). "Mobile health technology evaluation: the mHealth evidence workshop." *American journal of preventive medicine* **45**(2): 228-236.

Kurti, A. N. and J. Dallery (2013). "Internet-based contingency management increases walking in sedentary adults." *Journal of applied behavior analysis* **46**(3): 568-581.

Kwasnicka, D., S. U. Dombrowski, M. White and F. F. Sniehotta (2015). "Data-prompted interviews: Using individual ecological data to stimulate narratives and explore meanings." *Health Psychology* **34**(12): 1191.

L Tate, R., S. Mcdonald, M. Perdices, L. Togher, R. Schultz and S. Savage (2008). "Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the Single-Case Experimental Design (SCED) Scale." *Neuropsychological Rehabilitation* **18**(4): 385-401.

Lacroix, J., P. Saini and R. Holmes (2008). The relationship between goal difficulty and performance in the context of a physical activity intervention program. Proceedings of the 10th international conference on Human computer interaction with mobile devices and services, ACM.

Lane, J. D. and D. L. Gast (2014). "Visual analysis in single case experimental design studies: Brief review and guidelines." *Neuropsychological rehabilitation* **24**(3-4): 445-463.

Lane, N. D., E. Miluzzo, H. Lu, D. Peebles, T. Choudhury and A. T. Campbell (2010). "A survey of mobile phone sensing." *IEEE Communications magazine* **48**(9).

Lathia, N., V. Pejovic, K. K. Rachuri, C. Mascolo, M. Musolesi and P. J. Rentfrow (2013). "Smartphones for Large-Scale Behavior Change Interventions." *Pervasive Computing, IEEE* **12**(3): 66-73.

Lattie, E. G., Schueller, S. M., Sargent, E., Stiles-Shields, C., Tomasino, K. N., Corden, M. E., ... & Mohr, D. C. (2016). Uptake and usage of IntelliCare: a publicly available suite of mental health and well-being apps. *Internet interventions*, *4*, 152-158.

Ledford, J. R., J. D. Lane and K. E. Severini (2018). "Systematic use of visual analysis for assessing outcomes in single case design studies." *Brain Impairment* **19**(1): 4-17.

Ledger, D. and D. McCaffrey (2014). "Inside wearables: How the science of human behavior change offers the secret to long-term engagement." *Endeavour Partners* **200**(93): 1.

Lee, I.-M., E. J. Shiroma, F. Lobelo, P. Puska, S. N. Blair, P. T. Katzmarzyk and L. P. A. S. W. Group (2012). "Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy." *The lancet* **380**(9838): 219-229.

Lee, M.-H., S. Cha and T.-J. Nam (2015). Patina engraver: visualizing activity logs as patina in fashionable trackers. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM.

Lee, M. K., J. Kim, J. Forlizzi and S. Kiesler (2015). Personalization revisited: a reflective approach helps people better personalize health services and motivates them to increase physical activity. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Osaka, Japan, ACM**:** 743-754.

Levac, D., H. Colquhoun and K. K. O'Brien (2010). "Scoping studies: advancing the methodology." *Implementation Science* **5**(1): 69.

Leviton, L. C., L. K. Khan, D. Rog, N. Dawkins and D. Cotton (2010). "Evaluability assessment to improve public health policies, programs, and practices." *Annual review of public health* **31**.

Lewis, J. R. (1995). "IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use." *International Journal of Human-Computer Interaction* **7**(1): 57-78.

Lewis, T. L. and J. C. Wyatt (2014). "mHealth and mobile medical apps: a framework to assess risk and promote safer use." *Journal of medical Internet research* **16**(9).

Lewis, Z. H., E. J. Lyons, J. M. Jarvis and J. Baillargeon (2015). "Using an electronic activity monitor system as an intervention modality: a systematic review." *BMC public health* **15**(1): 585.

Liao, P., P. Klasnja, A. Tewari and S. A. Murphy (2016). "Sample size calculations for micro-randomized trials in mHealth." *Statistics in Medicine* **35**(12): 1944-1971.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., ... & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS medicine*, *6*(7), e1000100.

Lim, B. Y., A. Shick, C. Harrison, S. Hudson (2011). *Pediluma: Motivating Physical Activity Through Contextual Information and Social Influence*.

Lupton, D. (2014). "Critical Perspectives on Digital Health Technologies." *Sociology Compass* **8**(12): 1344-1359.

Lyon, A. R. and K. Koerner (2016). "User-Centered Design for Psychosocial Intervention Development and Implementation." *Clinical Psychology: Science and Practice* **23**(2): 180-200.

Lyons, E. J., Z. H. Lewis, B. G. Mayrsohn and J. L. Rowland (2014). "Behavior Change Techniques Implemented in Electronic Lifestyle Activity Monitors: A Systematic Content Analysis." *Journal of Medical Internet Research* **16**(8).

Maggin, D. M., H. Swaminathan, H. J. Rogers, B. V. O'keeffe, G. Sugai and R. H. Horner (2011). "A generalized least squares regression approach for computing effect sizes in single-case research: Application examples." *Journal of School Psychology* **49**(3): 301-321.

Maguire, M. (2001). "Methods to support human-centred design." *International journal of human-computer studies* **55**(4): 587-634.

Maitland, J., S. Sherwood, L. Barkhuus, I. Anderson, M. Hall, B. Brown, M. Chalmers and H. Muller (2006). Increasing the awareness of daily activity levels with pervasive computing. Pervasive Health Conference and Workshops, 2006, IEEE.

Maitland, J. and K. A. Siek (2009). Technological approaches to promoting physical activity. Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7, ACM.

Major, M. J. and M. Alford (2016). "Validity of the iPhone M7 motion co-processor as a pedometer for able-bodied ambulation." *Journal of sports sciences* **34**(23): 2160-2164.

Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: guided by information power. *Qualitative health research*, *26*(13), 1753-1760.

Manolov, R., V. Sierra, A. Solanas and J. Botella (2014). "Assessing Functional Relations in Single-Case Designs:Quantitative Proposals in the Context of the Evidence-Based Movement." *Behavior Modification* **38**(6): 878-913.

Marchand, E., E. Stice, P. Rohde and C. B. Becker (2011). "Moving from efficacy to effectiveness trials in prevention research." *Behaviour Research and Therapy* **49**(1): 32-41.

Marshall, M. N. (1996). "Sampling for qualitative research." *Family practice* **13**(6): 522-526.

Martin, S. S., D. I. Feldman, R. S. Blumenthal, S. R. Jones, W. S. Post, R. A. McKibben, E. D. Michos, C. E. Ndumele, E. V. Ratchford and J. Coresh (2015). "mActive: a randomized clinical trial of an automated mHealth intervention for physical activity promotion." *Journal of the American Heart Association* **4**(11): e002239.

Masters, K. (2014). "Health professionals as mobile content creators: Teaching medical students to develop mHealth applications." *Medical teacher* **36**(10): 883-889.

Matthews, C. E., B. E. Ainsworth, R. W. Thompson and J. D. Bassett (2002). "Sources of variance in daily physical activity levels as measured by an accelerometer." *Medicine and science in sports and exercise* **34**(8): 1376-1381.

Mattila, E., R. Lappalainen, J. Parkka, J. Salminen and I. Korhonen (2010). "Use of a mobile phone diary for observing weight management and related behaviours." *Journal of Telemedicine and Telecare* **16**(5): 260-264.

Mattila, E., J. Pärkkä, M. Hermersdorf, J. Kaasinen, J. Vainio, K. Samposalo, J. Merilahti, J. Kolari, M. Kulju and R. Lappalainen (2008). "Mobile diary for wellness management—results on usage and usability in two user studies." *IEEE Transactions on information technology in biomedicine* **12**(4): 501-512.

Maxwell, J. (1992). "Understanding and validity in qualitative research." *Harvard educational review* **62**(3): 279-301.

Maxwell, J. A. and D. M. Loomis (2003). "Mixed methods design: An alternative approach." *Handbook of mixed methods in social and behavioral research* **1**: 241-272.

Maxwell, J. A. and K. Mittapalli (2010). "Realism as a stance for mixed methods research." *Handbook of mixed methods in social & behavioral research*: 145-168.

May, C. R., Mair, F., Finch, T., MacFarlane, A., Dowrick, C., Treweek, S., ... & Murray, E. (2009). Development of a theory of implementation and integration: Normalization Process Theory. *Implementation Science*, *4*(1), 29.

McDonald, S., F. Quinn, R. Vieira, N. O'Brien, M. White, D. W. Johnston and F. F. Sniehotta (2017). "The state of the art and future opportunities for using longitudinal n-of-1 methods in health behaviour research: a systematic literature overview." *Health psychology review* **11**(4): 307-323.

McEvoy, P. and D. Richards (2006). "A critical realist rationale for using a combination of quantitative and qualitative methods." *Journal of Research in Nursing* **11**(1): 66-78.

McMahon, S. K., B. Lewis, J. M. Oakes, J. F. Wyman, W. Guan and A. J. Rothman (2017). "Assessing the Effects of Interpersonal and Intrapersonal Behavior Change Strategies on Physical Activity in Older Adults: a Factorial Experiment." *Annals of Behavioral Medicine*: 1-15.

McMillan, D., A. Morrison, O. Brown, M. Hall and M. Chalmers (2010). Further into the wild: Running worldwide trials of mobile systems. International Conference on Pervasive Computing, Springer.

McNamee, P., E. Murray, M. P. Kelly, L. Bojke, J. Chilcott, A. Fischer, R. West and L. Yardley (2016). "Designing and undertaking a health economics study of digital health interventions." *American journal of preventive medicine* **51**(5): 852-860.

Mechael, P. N. (2009). "The case for mHealth in developing countries." *Innovations: Technology, Governance, Globalization* **4**(1): 103-118.

Melton, B. F., M. P. Buman, R. L. Vogel, B. S. Harris and L. E. Bigham (2016). "Wearable Devices to Improve Physical Activity and Sleep: A Randomized Controlled Trial of College-Aged African American Women." *Journal of Black Studies* **47**(6): 610-625.

Mendis, S. (2014). *Global status report on noncommunicable diseases 2014*, World health organization.

Meredith, S. E., M. J. Grabinski and J. Dallery (2011). "Internet-based group contingency management to promote abstinence from cigarette smoking: A feasibility study." *Drug and Alcohol Dependence* **118**(1): 23-30.

Michie, S., C. Abraham, C. Whittington, J. McAteer and S. Gupta (2009). "Effective techniques in healthy eating and physical activity interventions: a meta-regression." *Health Psychology* **28**(6): 690.

Michie, S., S. Ashford, F. F. Sniehotta, S. U. Dombrowski, A. Bishop and D. P. French (2011). "A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the CALO-RE taxonomy." *Psychol Health* **26**(11): 1479-1498.

Michie, S., M. M. Van Stralen and R. West (2011). "The behaviour change wheel: a new method for characterising and designing behaviour change interventions." *Implementation science* **6**(1): 42.

Michie, S., West, R., Sheals, K., & Godinho, C. A. (2018). Evaluating the effectiveness of behavior change techniques in health-related behavior: a scoping review of methods used. *Translational behavioral medicine*, *8*(2), 212-224.

Michie, S., R. West, M. Johnston, P. Mac Aonghusa, J. Thomas, M. Kelly and J. Shawe-Taylor (2017). "The Human Behaviour Change Project: Digitising the knowledge base on effectiveness of behaviour change interventions." *Frontiers in Public Health*.

Michie, S., L. Yardley, R. West, K. Patrick and F. Greaves (2017). "Developing and Evaluating Digital Interventions to Promote Behavior Change in Health and Health Care: Recommendations Resulting From an International Workshop." *Journal of Medical Internet Research* **19**(6).

Middelweerd, A., J. S. Mollee, C. N. van der Wal, J. Brug and S. J. te Velde (2014). "Apps to promote physical activity among adults: a review and content analysis." *International Journal of Behavioral Nutrition and Physical Activity* **11**.

Miller, A. D. and E. D. Mynatt (2014). StepStream: a school-based pervasive social fitness system for everyday adolescent health. Proceedings of the 32nd annual ACM conference on Human factors in computing systems, ACM.

Mintel (2016). Brits step up to wearable technology. *Technology*. M. Press.

Miočević, M., D. P. MacKinnon and R. Levy (2017). "Power in Bayesian Mediation Analysis for Small Sample Research." *Structural Equation Modeling: A Multidisciplinary Journal*: 1-18.

Mohr, D. C., K. Cheung, S. M. Schueller, C. H. Brown and N. Duan (2013). "Continuous evaluation of evolving behavioral intervention technologies." *American journal of preventive medicine* **45**(4): 517-523.

Mohr, D. C., A. R. Lyon, E. G. Lattie, M. Reddy and S. M. Schueller (2017). "Accelerating digital mental health research from early design and creation to

successful implementation and sustainment." *Journal of medical Internet research* **19**(5).

Mohr, D. C., S. M. Schueller, W. T. Riley, C. H. Brown, P. Cuijpers, N. Duan, M. J. Kwasny, C. Stiles-Shields and K. Cheung (2015). "Trials of intervention principles: evaluation methods for evolving behavioral intervention technologies." *Journal of medical Internet research* **17**(7).

Mook, D. G. (1983). In defense of external invalidity. *American psychologist*, *38*(4), 379.

Moore, G. F., S. Audrey, M. Barker, L. Bond, C. Bonell, W. Hardeman, L. Moore, A. O'Cathain, T. Tinati and D. Wight (2015). "Process evaluation of complex interventions: Medical Research Council guidance." *bmj* **350**: h1258.

Morrison, A., Brown, O., McMillan, D., & Chalmers, M. (2011, May). Informed consent and users' attitudes to logging in large scale trials. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems* (pp. 1501-1506). ACM.

Morrison, A., D. McMillan, S. Reeves, S. Sherwood and M. Chalmers (2012). A hybrid mass participation approach to mobile software trials. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Austin, Texas, USA, ACM**:** 1311-1320.

Morrison, C. and G. Doherty (2014). "Analyzing engagement in a web-based intervention platform through visualizing log-data." *J Med Internet Res* **16**(11): e252.

Morrison, L. G. and C. Hargood, Xiaowen, S., Dennison, L., Joseph, J., Hughes, S., Michaelides, D.T., Johnston, D., Johnston, M., Michie, S., Little, P., Smith, P.W., Weal, M.J., Yardley, L. (2014). "Understanding usage of a hybrid website and smartphone app for weight management: a mixed-methods study." **16**(10): e201.

Morse, J. M. (2000). Determining sample size, Sage Publications Sage CA: Thousand Oaks, CA.

Munson, S. A. and S. Consolvo (2012). Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on.

Munson, S. A., E. Krupka, C. Richardson and P. Resnick (2015). Effects of public commitments and accountability in a technology-supported physical activity intervention. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM.

Muntaner, A., J. Vidal-Conti and P. Palou (2015). "Increasing physical activity through mobile device interventions: A systematic review." *Health Informatics Journal*.

Murphy, S. A. (2005). "An experimental design for the development of adaptive treatment strategies." *Statistics in medicine* **24**(10): 1455-1481.

Murray, E., E. B. Hekler, G. Andersson, L. M. Collins, A. Doherty, C. Hollis, D. E. Rivera, R. West and J. C. Wyatt (2016). Evaluating digital health interventions: key questions and approaches, Elsevier.

Murray, E., Z. Khadjesari, I. R. White, E. Kalaitzaki, C. Godfrey, J. McCambridge, S. G. Thompson and P. Wallace (2009). "Methodological challenges in online trials." *Journal of Medical Internet Research* **11**(2).

Nahum-Shani, I., E. B. Hekler and D. Spruijt-Metz (2015). "Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework." *Health psychology : official journal of the Division of Health Psychology, American Psychological Association* **34**(0): 1209-1219.

Nakajima, T. and V. Lehdonvirta (2013). "Designing motivation using persuasive ambient mirrors." *Personal and ubiquitous computing* **17**(1): 107-126.

Naughton, F., & Johnstone, D. (2014). "A starter kit for undertaking n-of-1 trials." *European Health Psychologist* **16**(5): 196-205.

Neil-Sztramko, S. E., C. C. Gotay, C. M. Sabiston, P. A. Demers and K. C. Campbell (2017). "Feasibility of a telephone and web-based physical activity intervention for women shift workers." *Translational Behavioral Medicine*: 1-9.

Nielsen (2011). State Of The Media. *The Mobile Media Report Q3*. Available at https://www.nielsen.com/us/en/insights/reports/2011/state-of-the-media--mobile-media-report-q3-2011.html [accessed 27/10/18]

Nilsen, W., S. Kumar, A. Shar, C. Varoquiers, T. Wiley, W. T. Riley, M. Pavel and A. A. Atienza (2012). "Advancing the science of mHealth." *Journal of health communication* **17**(sup1): 5-10.

Nock, M. K., B. D. Michel and V. I. Photos (2007). "Single-case research designs." *Handbook of research methods in abnormal and clinical psychology*: 337-350.

Oakley, A., V. Strange, C. Bonell, E. Allen, J. Stephenson and R. S. Team (2006). "Process evaluation in randomised controlled trials of complex interventions." *BMJ : British Medical Journal* **332**(7538): 413-416.

Ortiz, A. M., S. J. Tueller, S. L. Cook and R. D. Furberg (2016). "ActiviTeen: A Protocol for Deployment of a Consumer Wearable Device in an Academic Setting." *JMIR Research Protocols* **5**(3).

Pagliari, C. (2007). "Design and evaluation in eHealth: challenges and implications for an interdisciplinary field." *Journal of medical Internet research* **9**(2).

Parker, R. I. and K. J. Vannest (2012). "Bottom-up analysis of single-case research designs." *Journal of Behavioral Education* **21**(3): 254-265.

Patel, M. S., D. A. Asch, R. Rosin, D. S. Small, S. L. Bellamy, K. Eberbach, K. J. Walters, N. Haff, S. M. Lee, L. Wesby, K. Hoffer, D. Shuttleworth, D. H. Taylor, V. Hilbert, J. S. Zhu, L. Yang, X. M. Wang and K. G. Volpp (2016a). "Individual Versus Team-Based Financial Incentives to Increase Physical Activity: A Randomized, Controlled Trial." *Journal of General Internal Medicine* **31**(7): 746-754.

Patel, M. S., D. A. Asch, R. Rosin, D. S. Small, S. L. Bellamy, J. Heuer, S. Sproat, C. Hyson, N. Haff, S. M. Lee, L. Wesby, K. Hoffer, D. Shuttleworth, D. H. Taylor, V. Hilbert, J. S. Zhu, L. Yang, X. M. Wang and K. G. Volpp (2016b). "Framing Financial Incentives to Increase Physical Activity Among Overweight and Obese Adults A Randomized, Controlled Trial." *Annals of Internal Medicine* **164**(6): 385-+.

Patel, M. S., K. G. Volpp, R. Rosin, S. L. Bellamy, D. S. Small, M. A. Fletcher, R. Osman-Koss, J. L. Brady, N. Haff, S. M. Lee, L. Wesby, K. Hoffer, D. Shuttleworth, D. H. Taylor, V. Hilbert, J. S. Zhu, L. Yang, X. M. Wang and D. A. Asch (2016c). "A Randomized Trial of Social Comparison Feedback and Financial Incentives to Increase Physical Activity." *American Journal of Health Promotion* **30**(6): 416-424.

Patrick, K., E. B. Hekler, D. Estrin, D. C. Mohr, H. Riper, D. Crane, J. Godino and W. T. Riley (2016). "The Pace of Technologic Change." *American journal of preventive medicine* **51**(5): 816-824.

Paul, L., S. Wyke, S. Brewster, N. Sattar, J. M. R. Gill, G. Alexander, D. Rafferty, A. K. McFadyen, A. Ramsay and A. Dybus (2016). "Increasing physical activity in stroke survivors using STARFISH, an interactive mobile phone application: A pilot study." *Topics in Stroke Rehabilitation* **23**(3): 170-177.

Pejovic, V. and M. Musolesi (2014). Anticipatory mobile computing for behaviour change interventions. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. Seattle, Washington, ACM**:** 1025-1034.

Pellegrini, C. A., J. M. Duncan, A. C. Moller, J. Buscemi, A. Sularz, A. DeMott, A. Pictor, S. Pagoto, J. Siddique and B. Spring (2012). "A smartphone-supported weight loss program: design of the ENGAGED randomized controlled trial." *Bmc Public Health* **12**.

Pellegrini, C. A., J. Steglitz, W. Johnston, J. Warnick, T. Adams, H. G. McFadden, J. Siddique, D. Hedeker and B. Spring (2015). "Design and protocol of

a randomized multiple behavior change trial: Make Better Choices 2 (MBC2)." *Contemporary Clinical Trials* **41**: 85-92.

Pellegrini, C. A., S. D. Verba, A. D. Otto, D. L. Helsel, K. K. Davis and J. M. Jakicic (2012). "The comparison of a technology-based system and an in-person behavioral weight loss intervention." *Obesity* **20**(2): 356-363.

Penados, A. L., M. Gielen, P.-J. Stappers and T. Jongert (2010). "Get up and move: an interactive cuddly toy that stimulates physical activity." *Personal and Ubiquitous Computing* **14**(5): 397-406.

Perski, O., A. Blandford, R. West and S. Michie (2016). "Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis." *Translational Behavioral Medicine*: 1-14.

Perski, O., A. Blandford, R. West and S. Michie (2016). "Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis." *Translational behavioral medicine* **7**(2): 254-267.

Petticrew, M., L. Anderson, R. Elder, J. Grimshaw, D. Hopkins, R. Hahn, L. Krause, E. Kristjansson, S. Mercer and T. Sipe (2014). "Complex interventions and their implications for systematic reviews: a pragmatic approach." *Midirs Midwifery Digest* **24**(1): 15.

Petticrew, M., P. Tugwell, E. Kristjansson, S. Oliver, E. Ueffing and V. Welch (2011). "Damned if you do, damned if you don't: subgroup analysis and equity." *Journal of Epidemiology & Community Health*: jech. 2010.121095.

Pham, Q., D. Wiljer and J. A. Cafazzo (2016). "Beyond the Randomized Controlled Trial: A Review of Alternatives in mHealth Clinical Trial Methods." *JMIR mHealth and uHealth* **4**(3).

Pham, X. L., Nguyen, T. H., & Chen, G. D. (2017). Research Through the App Store: Understanding Participant Behavior on a Mobile English Learning App. *Journal of Educational Computing Research*

Pham, Q., Cafazzo, J. A., & Feifer, A. (2017). Adoption, acceptability, and effectiveness of a mobile health app for personalized prostate cancer survivorship care: protocol for a realist case study of the Ned App. *JMIR research protocols*, *6*(10).

Phillippi, J. and J. Lauderdale (2017). "A Guide to Field Notes for Qualitative Research: Context and Conversation." *Qualitative Health Research*: 1049732317697102.

Piwek, L., D. A. Ellis, S. Andrews and A. Joinson (2016). "The rise of consumer health wearables: promises and barriers." *PLoS Medicine* **13**(2): e1001953.

Ploderer, B., W. Smith, J. Pearce and R. Borland (2014). "A mobile app offering distractions and tips to cope with cigarette craving: a qualitative study." *JMIR mHealth and uHealth* **2**(2).

Polzien, K. M., Jakicic, J. M., Tate, D. F., & Otto, A. D. (2007). The efficacy of a technology-based system in a short-term behavioral weight loss intervention. *Obesity*, *15*(4), 825-830.

Poole, E. S., E. Eiríksdóttir, A. D. Miller, Y. Xu, R. Catrambone and E. D. Mynatt (2013). Designing for spectators and coaches: social support in pervasive health games for youth. Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Powell, R. A. and H. M. Single (1996). "Focus groups." *International journal for quality in health care* **8**(5): 499-504.

Price, M., E. K. Yuen, E. M. Goetter, J. D. Herbert, E. M. Forman, R. Acierno and K. J. Ruggiero (2014). "mHealth: a mechanism to deliver more accessible, more

effective mental health care." *Clinical psychology & psychotherapy* **21**(5): 427-436.

Proctor, E., H. Silmere, R. Raghavan, P. Hovmand, G. Aarons, A. Bunger, R. Griffey and M. Hensley (2011). "Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda." *Administration and Policy in Mental Health and Mental Health Services Research* **38**(2): 65-76.

Provost, F. and T. Fawcett (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*, " O'Reilly Media, Inc.".

Quesenbery, W. (2003). *The five dimensions of usability*, Lawrence Erlbaum Associates Mahwah, NJ.

Quintiliani, L. M., D. M. Mann, M. Puputti, E. Quinn and D. J. Bowen (2016). "Pilot and feasibility test of a mobile health-supported behavioral counseling intervention for weight management among breast cancer survivors." *JMIR cancer* **2**(1).

Rabbi, M., M. H. Aung, M. Zhang and T. Choudhury (2015). MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Osaka, Japan, ACM**:** 707-718.

Ragin, C. (1987). "The comparative method: Moving beyond qualitative and quantitative methods." *Berkeley: University of California*.

Read, J. C. (2008). "Validating the Fun Toolkit: an instrument for measuring children's opinions of technology." *Cognition, Technology & Work* **10**(2): 119-128.

Recio-Rodriguez, J. I., C. Martin-Cantera, N. Gonzalez-Viejo, A. Gomez-Arranz, M. S. Arietaleanizbeascoa, Y. Schmolling-Guinovart, J. A. Maderuelo-Fernandez, D. Perez-Arechaederra, E. Rodriguez-Sanchez, M. A. Gomez-Marcos and L. Garcia-Ortiz (2014). "Effectiveness of a smartphone application for improving

healthy lifestyles, a randomized clinical trial (EVIDENT II): study protocol." *BMC Public Health* **14**: 254.

Reid, R. E., J. A. Insogna, T. E. Carver, A. M. Comptour, N. A. Bewski, C. Sciortino and R. E. Andersen (2017). "Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment." *Journal of science and medicine in sport* **20**(6): 578-582.

Reijonsaari, K., A. Vehtari, O.-P. Kahilakoski, W. van Mechelen, T. Aro and S. Taimela (2012). "The effectiveness of physical activity monitoring and distance counseling in an occupational setting–Results from a randomized controlled trial (CoAct)." *BMC Public Health* **12**(1): 344.

Reilly, J. J., E. Methven, Z. C. McDowell, B. Hacking, D. Alexander, L. Stewart and C. J. Kelnar (2003). "Health consequences of obesity." *Archives of disease in childhood* **88**(9): 748-752.

Reiner, M., C. Niermann, D. Jekauc and A. Woll (2013). "Long-term health benefits of physical activity–a systematic review of longitudinal studies." *BMC public health* **13**(1): 813.

Ridgers, N. D., A. Timperio, H. Brown, K. Ball, S. Macfarlane, S. K. Lai, K. Richards, W. Ngan and J. Salmon (2017). "A cluster-randomised controlled trial to promote physical activity in adolescents: the Raising Awareness of Physical Activity (RAW-PA) Study." *Bmc Public Health* **17**.

Riley, W., R. Glasgow, L. Etheredge and A. Abernethy (2013). "Rapid, responsive, relevant (R3) research: a call for a rapid learning health research enterprise." *Clinical and Translational Medicine* **2**(1): 1-6.

Riley, W. T., R. E. Glasgow, L. Etheredge and A. P. Abernethy (2013). "Rapid, responsive, relevant (R3) research: a call for a rapid learning health research enterprise." *Clinical and translational medicine* **2**(1): 10.

Rippe, J. M. and S. Hess (1998). "The role of physical activity in the prevention and management of obesity." *Journal of the American Dietetic Association* **98**(10): S31-S38.

Robson, C. and K. McCartan (2016). *Real world research*, John Wiley & Sons.

Rodgers, M. M., V. M. Pai and R. S. Conroy (2015). "Recent Advances in Wearable Sensors for Health Monitoring." *Ieee Sensors Journal* **15**(6): 3119-3126.

Rogers, Y. and P. Marshall (2017). "Research in the Wild." *Synthesis Lectures on Human-Centered Informatics* **10**(3): i-97.

Rooksby, J., P. Asadzadeh, A. Morrison, C. McCallum, C. Gray and M. Chalmers (2016). Implementing ethics for a mobile app deployment. Proceedings of the 28th Australian Conference on Computer-Human Interaction, ACM.

Rooksby, J., M. Rost, A. Morrison and M. C. Chalmers (2014). Personal tracking as lived informatics. Proceedings of the 32nd annual ACM conference on Human factors in computing systems, ACM.

Rosenthal, R. (1966). "Experimenter effects in behavioral research."

Savovic J, Jones H, Altman D, et al. (2012) Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. *Health Technol Asses*s. 16(35):1–82.

Schlosser, R. W., D. L. Lee and O. Wendt (2008). "Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics." *Evidence-Based Communication Assessment and Intervention* **2**(3): 163-187.

Schoenfeld, W., W. Cumming and E. Hearst (1956). "On the classification of reinforcement schedules." *Proceedings of the national Academy of sciences* **42**(8): 563-570.

Schoeppe, S., S. Alley, W. Van Lippevelde, N. A. Bray, S. L. Williams, M. J. Duncan and C. Vandelanotte (2016). "Efficacy of interventions that use apps to improve diet, physical activity and sedentary behaviour: a systematic review." *International Journal of Behavioral Nutrition and Physical Activity* **13**(1): 127.

Schulz, K. F., D. G. Altman and D. Moher (2010). "CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials." *BMC medicine* **8**(1): 18.

Schwandt, T. and E. Halpern (1988). Linking auditing and metaevaluation: Enhancing quality in applied inquiry, Newbury Park, CA: Sage.

Schwandt, T. A. (1997). *Qualitative inquiry: A dictionary of terms*, Sage Publications, Inc.

Scruggs, T. E., M. A. Mastropieri and G. Casto (1987). "The quantitative synthesis of single-subject research: Methodology and validation." *Remedial and Special education* **8**(2): 24-33.

Sekhon, M., M. Cartwright and J. J. Francis (2017). "Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework." *BMC health services research* **17**(1): 88.

Sen, S. and C. B. Bhattacharya (2001). "Does doing good always lead to doing better? Consumer reactions to corporate social responsibility." *Journal of marketing Research* **38**(2): 225-243.

Shadish, W. R. (2014). "Statistical analyses of single-case designs: The shape of things to come." *Current Directions in Psychological Science* **23**(2): 139-146.

Shadish, W. R., T. D. Cook and D. T. Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*, Wadsworth Cengage learning.

Shadish, W. R., E. N. Kyse and D. M. Rindskopf (2013). "Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research." *Psychological methods* **18**(3): 385.

Shadish, W. R., A. F. Zuur and K. J. Sullivan (2014). "Using generalized additive (mixed) models to analyze single case designs." *Journal of school psychology* **52**(2): 149-178.

Shamseer, L., M. Sampson, C. Bukutu, C. H. Schmid, J. Nikles, R. Tate, B. C. Johnston, D. Zucker, W. R. Shadish and R. Kravitz (2015). "CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration." *Bmj* **350**: h1793.

Shiell, A., P. Hawe and L. Gold (2008). "Complex interventions or complex systems? Implications for health economic evaluation." *BMJ: British Medical Journal* **336**(7656): 1281.

Shiffman, S., A. A. Stone and M. R. Hufford (2008). "Ecological momentary assessment." *Annu. Rev. Clin. Psychol.* **4**: 1-32.

Shin, D. W., H.-K. Joh, J. M. Yun, H. T. Kwon, H. Lee, H. Min, J.-H. Shin, W. J. Chung, J. H. Park and B. Cho (2016). "Design and baseline characteristics of participants in the Enhancing Physical Activity and Reducing Obesity Through Smartcare and Financial Incentives (EPAROSFI): a pilot randomized controlled trial." *Contemporary clinical trials* **47**: 115-122.

Shiroma, E. J. and I.-M. Lee (2010). "Physical activity and cardiovascular health: lessons learned from epidemiological studies across age, gender, and race/ethnicity." *Circulation* **122**(7): 743-752.

Shuger, S. L., V. W. Barry, X. M. Sui, A. McClain, G. A. Hand, S. Wilcox, R. A. Meriwether, J. W. Hardin and S. N. Blair (2011). "Electronic feedback in a diet- and physical activity-based lifestyle intervention for weight loss: a randomized controlled trial." *International Journal of Behavioral Nutrition and Physical Activity* **8**.

Sieverink, F., S. M. Kelders, S. Akkersdijk, M. Poel, L. Tjin-Kam-Jet-Siemons and J. E. van Gemert-Pijnen (2016). Work in progress: a protocol for the collection, analysis, and interpretation of log data from eHealth technology. Fourth International Workshop on Behavior Change Support Systems, BCSS 2016: Epic for Change, the Pillars for Persuasive Technology for Smart Societies.

Sigal, R. J., G. P. Kenny, D. H. Wasserman, C. Castaneda-Sceppa and R. D. White (2006). "Physical Activity/Exercise and Type 2 Diabetes." *A consensus statement from the American Diabetes Association* **29**(6): 1433-1438.

Silverman, K., S. T. Higgins, R. K. Brooner, I. D. Montoya, E. J. Cone, C. R. Schuster and K. L. Preston (1996). "Sustained cocaine abstinence in methadone maintenance patients through voucher-based reinforcement therapy." *Archives of general psychiatry* **53**(5): 409-415.

Sim, J. (1998). "Collecting and analysing qualitative data: issues raised by the focus group." *Journal of advanced nursing* **28**(2): 345-352.

Slootmaker, S. M., M. J. Chinapaw, A. J. Schuit, J. C. Seidell and W. Van Mechelen (2009). "Feasibility and effectiveness of online physical activity advice based on a personal activity monitor: randomized controlled trial." *Journal of medical Internet research* **11**(3).

Slootmaker, S. M., M. J. C. A. Paw, A. J. Schuit, J. C. Seidell and W. Van Mechelen (2005). "Promoting physical activity using an activity monitor and a tailored web-based advice: design of a randomized controlled trial [ISRCTN93896459]." *BMC Public Health* **5**(1): 134.

Smith, J. D. (2012). "Single-case experimental designs: A systematic review of published research and current standards." *Psychological methods* **17**(4): 510.

Snodgrass, M. R., M. Y. Chung, H. Meadan and J. W. Halle (2018). "Social validity in single-case research: A systematic literature review of prevalence and application." *Research in developmental disabilities* **74**: 160-173.

Spencer, L. and J. Ritchie (2002). *Qualitative data analysis for applied policy research. Analyzing qualitative data*, Routledge**:** 187-208.

Statista. (2017). "Number of smartphone users worldwide from 2014 to 2020 (in billions)." *WebCite URL:* http://www.webcitation.org/6wvXEk8pWoriginal *URL:* https://www.statista.com/statistics/330695umber-of-smartphone-users-worldwide/ Retrieved 23/1/18.

Stawarz, K. and A. L. Cox (2015). Designing for health behavior change: HCI research alone is not enough. CHI'15 workshop: Crossing HCI and Health: Advancing Health and Wellness Technology Research in Home and Community Settings.

Stewart, A. L., K. M. Mills, A. C. King, W. L. Haskell, D. Gillis and P. L. Ritter (2001). "CHAMPS physical activity questionnaire for older adults: outcomes for interventions." *Medicine & Science in Sports & Exercise*.

Stuckey, M. I., S. W. Carter and E. Knight (2017). "The role of smartphones in encouraging physical activity in adults." *International journal of general medicine* **10**: 293.

Sucala, M., W. Nilsen and F. Muench (2017). "Building partnerships: a pilot study of stakeholders' attitudes on technology disruption in behavioral health delivery and research." *Translational behavioral medicine* **7**(4): 854-860.

Sugimoto, C. R., V. Larivière, C. Ni and B. Cronin (2013). "Journal acceptance rates: a cross-disciplinary analysis of variability and relationships with journal measures." *Journal of Informetrics* **7**(4): 897-906.

Tabak, M., H. Op den Akker and H. Hermens (2014a). "Motivational cues as real-time feedback for changing daily activity behavior of patients with COPD." *Patient Education and Counseling* **94**(3): 372-378.

Tabak, M., M, Vollenbroek-Hutten, P. van der Valk, J. van der Palen and H. J. Hermens (2014b). "A telerehabilitation intervention for patients with Chronic

Obstructive Pulmonary Disease: a randomized controlled pilot trial." *Clinical Rehabilitation* **28**(6): 582-591.

Tate, R. L., M. Perdices, U. Rosenkoetter, D. Wakim, K. Godbee, L. Togher and S. McDonald (2013). "Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale." *Neuropsychological rehabilitation* **23**(5): 619-638.

Taylor, D., J. Murphy, M. Ahmad, S. Purkayastha, S. Scholtz, R. Ramezani, I. Vlaev, A. Blakemore and A. Darzi (2016). "Quantified-Self for Obesity: Physical Activity Behaviour Sensing to Improve Health Outcomes." *Medicine Meets Virtual Reality 22: NextMed/MMVR22* **220**: 414.

Thomas, J. G. and D. S. Bond (2015). "Behavioral Response to a Just-in-Time Adaptive Intervention (JITAI) to Reduce Sedentary Behavior in Obese Adults: Implications for JITAI Optimization." *Health Psychology* **34**: 1261-1267.

Thompson, P. D., D. Buchner, I. L. Piña, G. J. Balady, M. A. Williams, B. H. Marcus, K. Berra, S. N. Blair, F. Costa, B. Franklin, G. F. Fletcher, N. F. Gordon, R. R. Pate, B. L. Rodriguez, A. K. Yancey and N. K. Wenger (2003). "Exercise and Physical Activity in the Prevention and Treatment of Atherosclerotic Cardiovascular Disease." *A Statement From the Council on Clinical Cardiology*. **107**(24): 3109-3116.

Thompson, W. G., C. L. Kuhle, G. A. Koepp, S. K. McCrady-Spitzer and J. A. Levine (2014). "“Go4Life” exercise counseling, accelerometer feedback, and activity levels in older people." *Archives of gerontology and geriatrics* **58**(3): 314-319.

Thompson, W. R. (2017). "Worldwide Survey Of Fitness Trends For 2018: The Crep Edition." *ACSM's Health & Fitness Journal* **21**(6): 10-19.

Thorndike, A. N., S. Mills, L. Sonnenberg, D. Palakshappa, T. Gao, C. T. Pau and S. Regan (2014). "Activity monitor intervention to promote physical activity of physicians-in-training: randomized controlled trial." *PLoS One* **9**(6): e100251.

Tomlinson, M., M. J. Rotheram-Borus, L. Swartz and A. C. Tsai (2013). "Scaling up mHealth: where is the evidence?" *PLoS medicine* **10**(2): e1001382.

Torous, J. and J. Firth (2016). "The digital placebo effect: mobile mental health meets clinical psychiatry." *The Lancet Psychiatry* **3**(2): 100-102.

Toval, A., B. Moros, J. Nicolas and J. Lasheras (2008). "Eight key issues for an effective reuse-based requirements process." *Computer Systems Science and Engineering* **23**(6): 373.

Tudor-Locke, C. and D. R. Bassett (2004). "How many steps/day are enough? Preliminary pedometer indices for public health." *Sports Medicine* **34**(1): 1-8.

Tudor-Locke, C., S. B. Sisson, T. Collova, S. M. Lee and P. D. Swan (2005). "Pedometer-determined step count guidelines for classifying walking intensity in a young ostensibly healthy population." *Canadian Journal of Applied Physiology-Revue Canadienne De Physiologie Appliquee* **30**(6): 666-676.

University of Southampton. (2016). "LifeGuide Community Website." Retrieved 25/10/2018, 2018, from https://www.lifeguideonline.org/.

Valbuena, D., B. G. Miller, A. L. Samaha and R. G. Miltenberger (2017). "Data presentation options to manage variability in physical activity research." *Journal of applied behavior analysis* **50**(3): 622-640.

Valentin, G. and A. M. Howard (2013). "Dealing with Childhood Obesity: Passive versus Active Activity Monitoring Approaches for Engaging Individuals in Exercise." *2013 Issnip Biosignals and Biorobotics Conference (Brc)*: 166-170.

Valle, C. G., A. M. Deal and D. F. Tate (2017). "Preventing weight gain in African American breast cancer survivors using smart scales and activity trackers: a randomized controlled pilot study." *Journal of Cancer Survivorship* **11**(1): 133-148.

van der Weegen, S., R. Verwey, M. Spreeuwenberg, H. Tange, T. van der Weijden and L. de Witte (2015). "It's LiFe! Mobile and Web-Based Monitoring and

Feedback Tool Embedded in Primary Care Increases Physical Activity: A Cluster Randomized Controlled Trial." *Journal of Medical Internet Research* **17**(7).

Van Gemert-Pijnen, J. E., S. M. Kelders and E. T. Bohlmeijer (2014). "Understanding the usage of content in a mental health intervention for depression: an analysis of log data." *J Med Internet Res* **16**(1): e27.

Van Hoye, K., F. Boen and J. Lefevre (2012). "The effects of physical activity feedback on behavior and awareness in employees: study protocol for a randomized controlled trial." *International journal of telemedicine and applications* **2012**: 10.

van Nassau, F., H. P. van der Ploeg, F. Abrahamsen, E. Andersen, A. S. Anderson, J. E. Bosmans, C. Bunn, M. Chalmers, C. Clissmann, J. M. R. Gill, C. M. Gray, K. Hunt, J. G. M. Jelsma, J. G. La Guardia, P. N. Lemyre, D. W. Loudon, L. Macaulay, D. J. Maxwell, A. McConnachie, A. Martin, N. Mourselas, N. Mutrie, R. Nijhuis-van der Sanden, K. O'Brien, H. V. Pereira, M. Philpott, G. C. Roberts, J. Rooksby, M. Rost, Ø. Røynesdal, N. Sattar, M. N. Silva, M. Sorensen, P. J. Teixeira, S. Treweek, T. van Achterberg, I. van de Glind, W. van Mechelen and S. Wyke (2016). "Study protocol of European Fans in Training (EuroFIT): a four-country randomised controlled trial of a lifestyle program for men delivered in elite football clubs." *BMC Public Health* **16**(1): 598.

Verwey, R., S. van der Weegen, M. Spreeuwenberg, H. Tange, T. van der Weijden and L. de Witte (2014a). "A monitoring and feedback tool embedded in a counselling protocol to increase physical activity of patients with COPD or type 2 diabetes in primary care: study protocol of a three-arm cluster randomised controlled trial." *Bmc Family Practice* **15**.

Verwey, R., S. van der Weegen, M. Spreeuwenberg, H. Tange, T. van der Weijden and L. de Witte (2014b). "A pilot study of a tool to stimulate physical activity in patients with COPD or type 2 diabetes in primary care." *Journal of Telemedicine and Telecare* **20**(1): 29-34.

Verwey, R., S. Van der Weegen, M. Spreeuwenberg, H. Tange, T. Van der Weijden and L. De Witte (2016). "Process evaluation of physical activity

counselling with and without the use of mobile technology: a mixed methods study." *International journal of nursing studies* **53**: 3-16.

Volkova, E., N. Li, E. Dunford, H. Eyles, M. Crino, J. Michie and C. N. Mhurchu (2016). "“Smart” RCTs: development of a smartphone app for fully automated nutrition-labeling intervention trials." *JMIR mHealth and uHealth* **4**(1).

Vorrink, S. N., H. S. Kort, T. Troosters, P. Zanen and J.-W. J. Lammers (2016). "Efficacy of an mHealth intervention to stimulate physical activity in COPD patients after pulmonary rehabilitation." *European Respiratory Journal*: ERJ-00083-02016.

Wadhwa, R., A. Chugh, A. Kumar, M. Singh, K. Yadav, S. Eswaran and T. Mukherjee (2015). *SenseX: Design and Deployment of a Pervasive Wellness Monitoring Platform for Workplaces*. *Service-Oriented Computing*. A. Barros, D. Grigori, N. C. Narendra and H. K. Dam. 9435**:** 427-443.

Walker, M., L. Takayama and J. A. Landay (2002). High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. Proceedings of the human factors and ergonomics society annual meeting, SAGE Publications Sage CA: Los Angeles, CA.

Walsh, G. and J. Golbeck (2014). StepCity: a preliminary investigation of a personal informatics-based social game on behavior change. CHI'14 Extended Abstracts on Human Factors in Computing Systems, ACM.

Walsh, J. C., T. Corbett, M. Hogan, J. Duggan and A. McNamara (2016). "An mHealth Intervention Using a Smartphone App to Increase Walking Behavior in Young Adults: A Pilot Study." *Jmir Mhealth and Uhealth* **4**(3).

Walters, D. L., A. Sarela, A. Fairfull, K. Neighbour, C. Cowen, B. Stephens, T. Sellwood, B. Sellwood, M. Steer, M. Aust, R. Francis, C. K. Lee, S. Hoffman, G. Brealey and M. Karunanithi (2010). "A mobile phone-based care model for outpatient cardiac rehabilitation: the care assessment platform (CAP)." *Bmc Cardiovascular Disorders* **10**.

Wang, J. B., L. A. Cadmus-Bertram, L. Natarajan, M. M. White, H. Madanat, J. F. Nichols, G. X. Ayala and J. P. Pierce (2015). "Wearable sensor/device (Fitbit One) and SMS text-messaging prompts to increase physical activity in overweight and obese adults: a randomized controlled trial." *Telemedicine and e-Health* **21**(10): 782-792.

Ward-Horner, J. and P. Sturmey (2010). "Component analyses using single-subject experimental designs: A review." *Journal of applied behavior analysis* **43**(4): 685-704.

Watson, P. and E. A. Workman (1981). "The non-concurrent multiple baseline across-individuals design: An extension of the traditional multiple baseline design." *Journal of Behavior Therapy and Experimental Psychiatry* **12**(3): 257-259.

Watson, S., J. V. Woodside, L. J. Ware, S. J. Hunter, A. McGrath, C. R. Cardwell, K. M. Appleton, I. S. Young and M. C. McKinley (2015). "Effect of a web-based behavior change program on weight loss and cardiovascular risk factors in overweight and obese adults at high risk of developing cardiovascular disease: Randomized controlled trial." *Journal of Medical Internet Research* **17**(7).

Weber, D., A. Voit, T. Dingler, M. Kallert and N. Henze (2016). Assessment of an unobtrusive persuasive system for behavior change in home environments. *Proceedings of the 5th ACM International Symposium on Pervasive Displays*. Oulu, Finland, ACM**:** 245-246.

Webster, J. and R. T. Watson (2002). "Analyzing the past to prepare for the future: Writing a literature review." *MIS quarterly*: xiii-xxiii.

Webb, J., Foster, J., & Poulter, E. (2016). Increasing the frequency of physical activity very brief advice for cancer patients. Development of an intervention using the behaviour change wheel. *public health*, *133*, 45-56.

White, O. R. and N. G. Haring (1976). *Exceptional teaching: A multimedia training package*, Merrill Publishing Company.

Whittaker, R. (2012). "Issues in mHealth: findings from key informant interviews." *Journal of medical Internet research* **14**(5).

Whittaker, R., S. Merry, E. Dorey and R. Maddison (2012). "A development and evaluation process for mHealth interventions: examples from New Zealand." *Journal of health communication* **17**(sup1): 11-21.

Wilson, J. and D. Rosenberg (1988). *Rapid prototyping for user interface design. Handbook of human-computer interaction*, Elsevier**:** 859-875.

Winter, S. J., J. L. Sheats and A. C. King (2016). "The use of behavior change techniques and theory in technologies for cardiovascular disease prevention and treatment in adults: a comprehensive review." *Progress in cardiovascular diseases* **58**(6): 605-612.

Wolery, M. (2013). "A commentary: Single-case design technical document of the What Works Clearinghouse." *Remedial and Special Education* **34**(1): 39-43.

Wolf, M. M. (1978). "Social validity: the case for subjective measurement or how applied behavior analysis is finding its heart." *Journal of Applied Behavior Analysis* **11**(2): 203.

Wolk, S., T. Meißner, S. Linke, B. Müssle, A. Wierick, A. Bogner, D. Sturm, N. N. Rahbari, M. Distler and J. Weitz (2017). "Use of activity tracking in major visceral surgery—the Enhanced Perioperative Mobilization (EPM) trial: study protocol for a randomized controlled trial." *Trials* **18**(1): 77.

Wyrick, D. L., K. L. Rulison, M. Fearnow-Kenney, J. J. Milroy and L. M. Collins (2014). "Moving beyond the treatment package approach to developing behavioral interventions: addressing questions that arose during an application of the Multiphase Optimization Strategy (MOST)." *Translational behavioral medicine* **4**(3): 252-259.

Xu, Y., E. S. Poole, A. D. Miller, E. Eiriksdottir, R. Catrambone and E. D. Mynatt (2012). Designing pervasive health games for sustainability, adaptability and

sociability. Proceedings of the International Conference on the Foundations of Digital Games, ACM.

Yardley, L., B. J. Spring, H. Riper, L. G. Morrison, D. H. Crane, K. Curtis, G. C. Merchant, F. Naughton and A. Blandford (2016). "Understanding and promoting effective engagement with digital behavior change interventions." *American Journal of Preventive Medicine* **51**(5): 833-842.

Yin, R. (1984). Applied Social Research Methods. Vol. 5. Case Study Research, Beverly Hills, CA: Sage.

Yingling, L. R., A. T. Brooks, G. R. Wallen, M. Peters-Lawrence, M. McClurkin, R. Cooper-McCann, K. L. Wiley, V. Mitchell, J. N. Saygbe, T. D. Johnson, K. E. Curry, A. A. Johnson, A. P. Graham, L. A. Graham and T. M. Powell-Wiley (2016). "Community Engagement to Optimize the Use of Web-Based and Wearable Technology in a Cardiovascular Health and Needs Assessment Study: A Mixed Methods Approach." *Jmir Mhealth and Uhealth* **4**(2): 38-55.

Yuan, Y. and R. H. Hunt (2009). "Systematic reviews: the good, the bad, and the ugly." *The American journal of gastroenterology* **104**(5): 1086.

Ziebland, S. and A. McPherson (2006). "Making sense of qualitative data analysis: an introduction with illustrations from DIPEx (personal experiences of health and illness)." *Medical education* **40**(5): 405-414.

Zuckerman, O. and A. Gal-Oz (2014). "Deconstructing gamification: evaluating the effectiveness of continuous measurement, virtual rewards, and social comparison for promoting physical activity." *Personal and Ubiquitous Computing* **18**(7): 1705-1719.